An International ICT Journal
Featuring Industry-University-Institute Cooperation

ISSN 1673-5188
CN 34-1294/TN

ZTE COMMUNICATIONS

中兴通讯技术（英文版）

http://zte.magtechjournal.com

December 2023, Vol. 21 No. 4

Special Topic:
3D Point Cloud Processing and Applications

ISSN 1673-5188

CONTENTS

ZTE COMMUNICATIONS
December 2023 Vol. 21 No. 4 (Issue 85, End of Volume)

Special Topic ▶

3D Point Cloud Processing and Applications

Research Papers ▶

Roundup ▶

Serial parameters: CN 34-1294/TN*2003*q*16*106*en*P*¥20.00*2200*12*2023-12

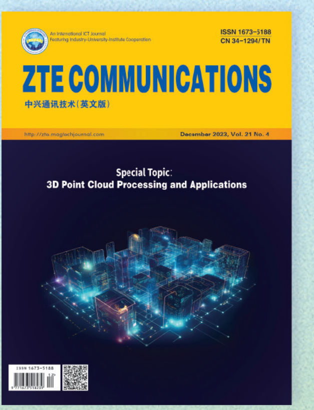
Statement

ZTE COMMUNICATIONS

ZTE Communications is a peer-reviewed international ICT journal (CN 34-1294/TN and ISSN 1673-5188) featuring industry-university-institute cooperation. It is published quarterly as a printed publication and can also be freely accessed at https://zte.magtechjournal.com.

ZTE Communications was founded in 2003 and has a readership of more than 100 000. The English version is distributed to universities, colleges, and research institutes in more than 100 countries. The journal covers a wide range of topics in the field of ICT. The Editorial Board members are distinguished domestic and international experts. The journal has become an integrated forum for university academics and industry researchers from around the world.

The journal uses ScholarOne Manuscripts, a professional web-based manuscript submission and peer-review tracking system. Authors must submit manuscripts electronically to https://mc03.manuscriptcentral.com/ztecom.



2024 Special Topics in *ZTE Communications*

Issue 1: Near-Field Communication and Sensing Towards 6G
WEI Guo, University of Science and Technology of China, China
CHEN Li, University of Science and Technology of China, China
ZHAO Yajun, ZTE Corporation, China

Issue 2: Advancements in Web3 Infrastructure for the Metaverse
Victor C. M. LEUNG, Shenzhen University, China
CAI Wei, CUHK-Shenzhen, China

Issue 3: Integrated Sensing and Communication (ISAC) Technologies for Future Wireless Systems
YUAN Jinhong, University of New South Wales, Australia

Issue 4: Optoelectronic Integrated Chips, Systems, and Key Technologies
WANG Yongjin, Nanjing University of Posts and Telecommunications, China

Guest Editorial >>>

Special Topic on 3D Point Cloud Processing and Applications

Guest Editors

SUN Huifang

LI Ge

CHEN Siheng

LI Li

GAO Wei

3D point cloud processing has redefined the way we perceive and interact with digital spatial data. By translating physical entities into a collection of 3D points, it offers an accurate digital model of our surroundings. This emerging field of 3D point-based representation has piqued interest significantly over recent years, owing to its capacity to depict detailed spatial environments, thereby bridging the gap between virtual and real dimensions. Numerous applications, including virtual reality, augmented reality, and advanced mapping, have greatly benefited from this technology, allowing for immersive experiences and accurate spatial analysis. However, the journey from raw spatial data to refined point cloud representations is fraught with challenges, including storage and computational demands, noise handling and the quest for efficient compression techniques.

In this special issue on 3D point cloud processing and applications, we present a curated series of articles that dive deep into these challenges, suggesting innovative strategies and methodologies tailored to address them. The selected contributions touch upon a diverse spectrum of topics within the realm of point cloud processing. They discuss novel compression algorithms, delve into quality assessment metrics, elucidate advanced rendering techniques, and highlight the nuances of feature extraction, among other pivotal areas. The call for papers for this special issue attracted excellent submissions, indicating the growing significance of this field. Following rigorous reviews, we are proud to present six standout papers that not only showcase cutting-edge research but also set the direction for future endeavors in this domain.

The first paper titled "Perceptual Quality Assessment for Point Clouds: A Survey" delivers a comprehensive overview of how the visual quality of point clouds is gauged. Traditional quality assessment methods fall short when applied to point cloud data. This survey presents the significance of point cloud quality assessment, discussing common distortions, experimental setups, and subjective databases. It contrasts model-based and projection-based objective methods, and the performance of these methods across various databases is analyzed. Experimental insights underline the utility and efficacy of the presented methods.

The second paper titled "Spatio-Temporal Context-Guided Algorithm for Lossless Point Cloud Geometry Compression" addresses the challenges faced during the compression of point cloud data. Traditional compression techniques struggle with the irregular distribution of point cloud data in space and time. This paper introduces an innovative context-guided algorithm that slices point clouds and employs the travelling salesman algorithm to predict compression. Testing results emphasize its robustness, presenting a feasible avenue for efficient 3D point cloud compression (PCC).

The third paper titled "Lossy Point Cloud Attribute Compression with Subnode-Based Prediction" shines light on the advances in 3D point cloud compression. With the Moving Picture Expert Group (MPEG) working towards a standard for PCC, the paper highlights the challenges in current attribute compression techniques. It introduces a subnode-based prediction method, leveraging spatial relationships for improved

DOI:10.12142/ZTECOM.202304001
Citation (Format 1): SUN H F, LI G, CHEN S H, et al. Special topic on 3D point cloud processing and applications [J]. ZTE Communications, 2023, 21(4): 1–2. DOI: 10.12142/ZTECOM.202304001
Citation (Format 2): H. F. Sun, G. Li, S. H. Chen, et al., "Special topic on 3D point cloud processing and applications," *ZTE Communications*, vol. 21, no. 4, pp. 1–2, Dec. 2023. doi: 10.12142/ZTECOM.202304001.

precision. Experimental results showcase its superior performance over existing MPEG standards.

The fourth paper titled "Point Cloud Processing Methods for 3D Point Cloud Detection Tasks" revolves around the pivotal role of 3D point cloud processing in object detection. Given the complexity of data acquired from LiDAR sensors, the paper offers a review of point cloud processing methods and how they influence detection outcomes. The discussion underscores the evolution of voxelization and sampling strategies, emphasizing their implications for feature extraction and final detection performance.

The fifth paper titled "Perceptual Optimization for Point-Based Point Cloud Rendering" delves into the challenges in point-based rendering for point clouds. The established method of determining rendering radius using neighboring points' distances is problematic. The paper introduces an outlier detection mechanism that optimizes the perceptual quality of rendering, using local and global geometric features to detect outliers. Results confirm the significant improvements in rendering quality with this approach.

The sixth paper titled "Local Scenario Perception and Web AR Navigation" explores the exciting convergence of web technologies and augmented reality (Web AR). As Web AR grapples with computational demands, the paper introduces an indoor navigation system based on local point cloud map positioning. This novel approach minimizes the need for external sensors, highlighting a promising avenue for precise and widespread application of Web AR navigation.

To conclude, this special issue aims to be an indispensable guide for researchers, industry experts, and students delving into 3D point cloud processing and its varied applications. We anticipate that the content will spur more research and advancements, shaping the future trajectory of digital spatial data analysis. Our deepest gratitude extends to all the authors, reviewers, and editorial staff for their invaluable contributions that have made this issue a success. We earnestly hope that the articles in this special issue offer both clarity and insight to all readers in this emerging domain.

Biographies

SUN Huifang graduated from Harbin Engineering University, China, and received his PhD degree from University of Ottawa, Canada. He was an associate professor in Fairleigh Dickinson University in 1990. He joined Sarnoff Corporation in 1990 as a member of technical staff and was promoted to a technology leader for digital video communication. In 1995, he joined Mitsubishi Electric Research Laboratories (MERL) and was promoted as Vice President and Deputy Director in 2003 and currently is a retired Fellow of MERL. He has co-authored two books and published more than 160 journal and conference papers. He holds 64 US patents. He won the Technical Achievement Award for optimization and specification of the Grand Alliance HDTV video compression algorithm in 1994 at Sarnoff Lab. He received the best paper award of 1992 *IEEE Transaction on Consumer Electronics*, the best paper award of 1996 *ICCE* and the best paper award of 2003 *IEEE Transaction on CSVT*. He was an associate editor for *IEEE Transaction on Circuits and Systems for Video Technology* and was the Chair of Visual Processing Technical Committee of IEEE Circuits and System Society. He is an IEEE life fellow.

LI Ge is currently a full professor at the School of Electronic and Computer Engineering, Peking University, China. He received his PhD degree from the Department of Electrical Engineering, Auburn University, USA in 1999. He has several years of research work experience in industry. His research interests include point cloud compression and its standardization, image/video processing and analysis, machine learning, and signal processing. He has published over 100 high quality papers and holds lots of granted US and global patents. He actively submitted many technical proposals to MPEG PCC and is also currently the Lead Chair for the standardization of point cloud compression in the Audio Video Coding Standard (AVS) Workgroup of China. He served as the Panel Chair of IEEE ICME 2021, the International Liaison Chair of PCS 2022, etc.

CHEN Siheng is a tenure-track associate professor of Shanghai Jiao Tong University, China. Before joining Shanghai Jiao Tong University, he was a research scientist at Mitsubishi Electric Research Laboratories (MERL) and an autonomy engineer at Uber Advanced Technologies Group (ATG), working on the perception and prediction systems of self-driving cars. Before joining industry, Dr. CHEN was a postdoctoral research associate at Carnegie Mellon University, USA. He received his doctorate in electrical and computer engineering from Carnegie Mellon University. He has published over 80 papers on prestigious venues, including *Nature Computational Science*, *Nature Scientific Data*, *IEEE Signal Processing Magazine*, *IEEE Transactions on Signal Processing*, *IEEE Transactions on Image Processing*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, ICASSP, NeurIPS, ICML, ICLR and CVPR. His work on sampling theory of graph data received the 2018 IEEE Signal Processing Society Young Author Best Paper Award. His co-authored paper on structural health monitoring received ASME SHM/NDE 2020 Best Journal Paper Runner-Up Award and another paper on 3D point cloud processing received the Best Student Paper Award at 2018 IEEE Global Conference on Signal and Information Processing. Dr. CHEN contributed to the project of scene-aware interaction, winning MERL President's Award. He serves as the associate editor for *IEEE Transactions on Signal and Information Processing over Networks*. His research interests include collaborative AI and 3D scene understanding.

LI Li received his BS and PhD degrees in electronic engineering from University of Science and Technology of China (USTC), China in 2011 and 2016, respectively. He was a visiting assistant professor in University of Missouri-Kansas City, USA from 2016 to 2020. He joined the Department of Electronic Engineering and Information Science of USTC as a research fellow in 2020 and became a professor in 2022. His research interests include image/video/point cloud coding and processing. He received the Best 10% Paper Award at the 2016 IEEE Visual Communications and Image Processing (VCIP) and the 2019 IEEE International Conference on Image Processing (ICIP).

GAO Wei is an assistant professor at the School of Electronic and Computer Engineering, Peking University, Shenzhen, China. His research has been focused on perception-inspired multimedia coding and processing, including both efficient algorithms and systems. He won the 2021 IEEE Multimedia Rising Star Runner Up Award for Outstanding Early-stage Career Achievements in the area of 3D Immersive media research. He is currently serving or has served as the associate editor for several international journals on multimedia computing and machine leaning, including *Signal Processing*, etc. He is also an Elected Member of IEEE CASS VSPC-TC, and APSIPA IVM-TC. He has organized workshops and special sessions at ICME 2023, ACM MM 2022, VCIP 2022, and ICME 2021. He is a tutorial speaker for point cloud related topics at ICME 2023 and ICIP 2023. He is also the leader to establish several open source projects, including OpenPointCloud, OpenAICoding, etc. He is a senior member of IEEE.

Perceptual Quality Assessment for Point Clouds : A Survey

ZHOU Yingjie, ZHANG Zicheng, SUN Wei,
MIN Xiongkuo, ZHAI Guangtao
(Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai 200240, China)

DOI: 10.12142/ZTECOM.202304002

https: //kns.cnki.net/kcms/detail/34.1294.TN.20231120.1548.002.html, published online November 21, 2023

Manuscript received: 2023-10-07

Abstract: A point cloud is considered a promising 3D representation that has achieved wide applications in several fields. However, quality degradation inevitably occurs during its acquisition and generation, communication and transmission, and rendering and display. Therefore, how to accurately perceive the visual quality of point clouds is a meaningful topic. In this survey, we first introduce the point cloud to emphasize the importance of point cloud quality assessment (PCQA). A review of subjective PCQA is followed, including common point cloud distortions, subjective experimental setups and subjective databases. Then we review and compare objective PCQA methods in terms of model-based and projection-based. Finally, we provide evaluation criteria for objective PCQA methods and compare the performances of various methods across multiple databases. This survey provides an overview of classical methods and recent advances in PCQA.

Keywords: point cloud quality assessment; PCQA databases; subjective quality assessment; objective quality assessment

Citation (Format 1): ZHOU Y J, ZHANG Z C, SUN W, et al. Perceptual quality assessment for point clouds : a survey [J]. ZTE Communications, 2023, 21(4): 3 – 16. DOI: 10.12142/ZTECOM.202304002

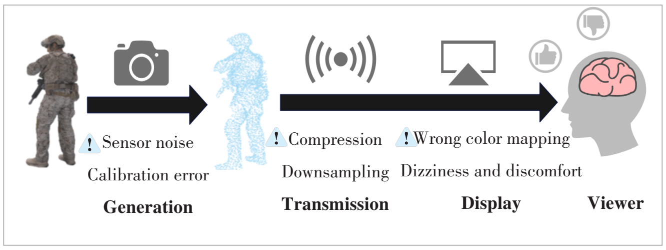
Citation (Format 2): Y. J. Zhou, Z. C. Zhang, W. Sun, et al., "Perceptual quality assessment for point clouds : a survey," *ZTE Communications*, vol. 21, no. 4, pp. 3 – 16, Dec. 2023. doi: 10.12142/ZTECOM.202304002.

1 Introduction

In recent years, three-dimensional (3D) data have garnered increasing attention due to its indispensable role in various applications, ranging from computer graphics and computer-aided design to autonomous navigation, augmented reality, and environmental modeling. Point clouds, as one of the fundamental representations of 3D data, have emerged as a prominent data structure capable of capturing the spatial information of objects and scenes with unparalleled fidelity. A point cloud is a collection of points in a 3D coordinate system, where each point represents a precise location in space, often obtained through laser scanning, photogrammetry, or other 3D sensing techniques. Previous research[1–3] has demonstrated the inestimable value of such data-rich structures in generating accurate 3D models, facilitating object recognition, and enabling realistic visual simulations.

While the adoption of point clouds has led to groundbreaking advancements in numerous fields, ensuring the quality and fidelity of these data representations poses significant challenges. As shown in Fig. 1, point clouds are not immune to quality degradation, particularly during the process of generation and transmission. The transfer of point clouds across networks, storage systems, or different software applications can introduce various forms of distortion, noise, and loss of information, which may substantially impact the utility and accuracy of the 3D models they represent. Therefore, point cloud quality assessment (PCQA) is of great importance in the research and development process. The need to assess the fidelity, integrity, and reliability of point clouds is becoming increasingly evident to preserve the quality of 3D data in practical applications. The availability of high-quality point clouds is fundamental to the success of downstream tasks (e. g., 3D reconstruction, object recognition, and analysis), as well as to the overall efficiency and robustness of various 3D-based systems.

Based on these considerations, this paper aims to provide insights into how point cloud quality can be comprehensively assessed. By carefully reviewing and summarizing existing methods and research results, we provide an overview of the current developments in the field of PCQA. In addition, this



▲Figure 1. Illustration of how distortions are generated in point clouds paper aims to promote the continuous progress and practical application of point cloud technology. A deeper understanding of point cloud quality assessment can pave the way for improved 3D data utilization in various industries and open up new possibilities in the fields of design, analytics, and immersive experiences.

2 Subjective Point Cloud Quality Assessment

This chapter delves into the assessment of human perception of point clouds with the aim of understanding how users subjectively perceive the quality of these 3D data representations. The chapter begins by describing various common types of point cloud distortion that may affect human perception. By shedding light on these distortions, the chapter highlights the importance of addressing these issues to improve the overall quality of point cloud visualization and ensure a more accurate user experience. An introduction to existing point cloud databases follows the discussion of point cloud distortion. This section focuses on databases curated specifically for subjective quality assessment purposes. The significance of such databases lies in their ability to provide researchers with carefully controlled test cases that allow for the systematic study of human perception at different levels of distortion.

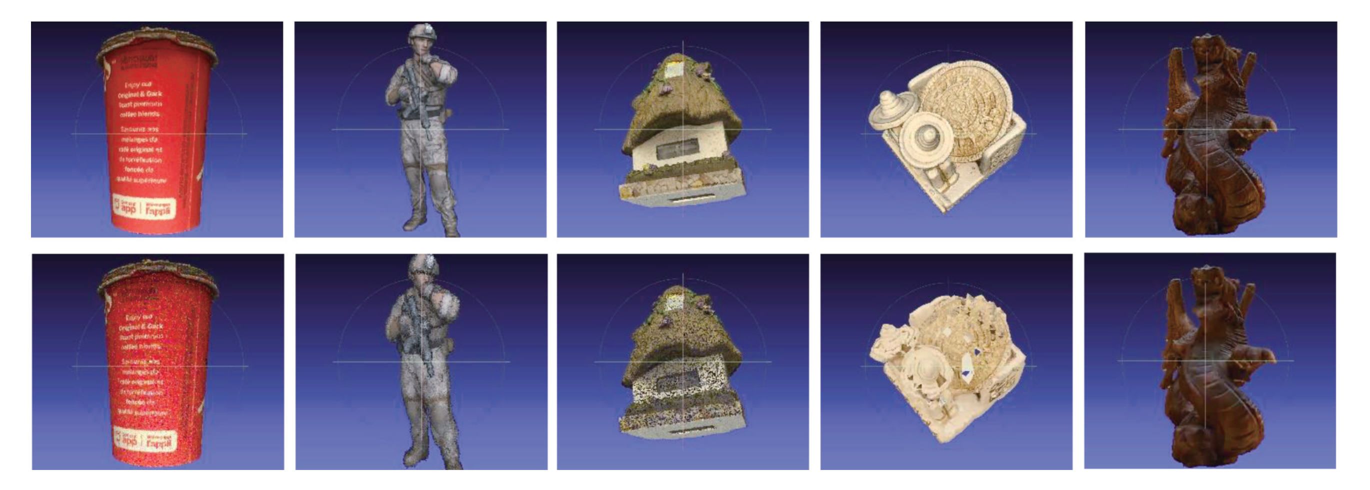
2.1 Common Types of Distortion in Point Clouds

As the point cloud distortion modeled in LS-PCQA[4] (a large-scale PCQA dataset) by LIU et al., point clouds are not only subjected to various degrees of noise, compression, and sampling, but even localized distortions such as loss, rotation, and Audio Video Coding Standard (AVS) during the actual generation and communication transmission process. We select four most common types of point cloud distortion to introduce, including color noise, geometric noise, downsampling and point cloud compression. The visual effects of the various distortions are shown in Fig. 2.

1) Color noise. Point cloud color noise is defined as unwanted variations and inaccuracies in the color information associated with individual points in a 3D point cloud. When acquiring point cloud data from various sources such as 3D scanners and LiDAR systems, color information is typically captured along with the 3D coordinates of each point. However, due to factors such as sensor noise, lighting conditions, and calibration errors, the color values assigned to the points may deviate from the true color of the corresponding object or surface in the real world. This may result in an inconsistent visual appearance of the point cloud and affect subsequent applications that rely on accurate color information.

2) Geometry noise. Point cloud geometry noise is the inherent irregularities and inaccuracies in the spatial coordinates of individual points in a 3D point cloud. These inaccuracies can arise from a variety of causes, including sensor limitations, measurement errors, calibration issues and occlusions during data acquisition. As a result, the point cloud may contain shifted or misaligned points, resulting in reduced geometric accuracy and fidelity. Geometry noise can adversely affect the quality of the 3D model derived from the point cloud as well as subsequent tasks.

3) Downsampling. Point cloud downsampling is a key technique used to reduce the data size of 3D point clouds while preserving the underlying structural and spatial information. Large-scale point clouds acquired from 3D scanning or LiDAR systems may contain millions or billions of points, and processing and storing these points require extensive computation. Downsampling involves the systematic removal of a subset of points from the original data, thereby effectively simplifying its representation without significantly affecting its overall shape and characteristics. However, downsampling also



▲Figure 2. Visualization of common point cloud distortions. The first row shows the reference point clouds and the second row from left to right shows the distortion effects of color noise, geometric noise, downsampling, geometry-based point cloud compression (GPCC) and video-based point cloud compression (VPCC), respectively. The selected point clouds shown are derived from existing databases[5–6] or related research[7] poses challenges, as indiscriminate removal of points might compromise fine details and key features, which could affect downstream applications like object recognition or surface reconstruction.

4) Compression. Point cloud compression is a vital area of research that aims to reduce the storage and transmission requirements of 3D point cloud data while preserving its essential geometric and semantic properties. One prominent approach to point cloud compression is the development of the video-based point cloud compression (VPCC) standard, which leverages the coding efficiency of video compression techniques adapted to the point cloud domain. VPCC efficiently represents point clouds by exploiting temporal redundancies and inter-frame dependencies, enabling high compression ratios while maintaining visual quality and accuracy. Another significant advancement in point cloud compression is the geometry-based point cloud compression (GPCC) standard. GPCC focuses on the efficient compression of point cloud geometry, utilizing various techniques like octree-based coding, predictive coding, and attribute coding. By considering the geometric properties of point clouds, GPCC achieves superior compression performance while facilitating fast and reliable decompression for real-time applications. VPCC and GPCC play a key role in optimizing the storage and delivery of massive point cloud databases, making them more accessible and usable in a variety of applications such as virtual reality, augmented reality, and cloud-based services. However, at the same time, point cloud compression inevitably leads to degradation of point cloud quality.

2.2 Common Subjective Experimental Setups

Presenting point cloud content is essential to harness the valuable information it contains. As a versatile data format, point clouds can be visualized using various methods. Traditional 2D monitors allow for a flat, easily accessible representation of point clouds, enabling researchers and users to explore the data from different angles. On the other hand, 3D monitors provide a more immersive experience, allowing a deeper understanding of the spatial relationships among the points. Furthermore, head-mounted devices (HMDs), such as virtual reality (VR) headsets, take the presentation of point clouds to another level, offering an unparalleled sense of presence and interaction with the 3D data. The summary of existing point cloud subjective evaluation works is presented in Table 1, with significant differences in interaction methods, viewing displays, scoring methods, and more. There are three prevalent scoring methodologies employed in the assessment of perceptual quality: the Double-Stimulus Impairment Scale (DSIS), Absolute Category Rating (ACR), and Pairwise Comparison (PWC). In DSIS, evaluators are presented with a reference point cloud and a distorted point cloud, with the task of rating the quality of the distorted point cloud. Conversely, in ACR, evaluators are tasked with categorizing each distorted point cloud into predefined quality categories, such as "excellent," "good," "fair," or "poor." In the case of PWC, evaluators are presented with pairs of distorted point clouds and are required to indicate which of the two exhibits superior quality.

▼ Table 1. Summary of the experimental setups for subjective cloud quality assessment

Related Work	Display	Interaction	Methodology
Work of ALEXIOU et al.[14]	2D monitor	×	DSIS
Work of ALEXIOU and EBRAHIMI[13]	2D monitor	×	DSIS
Work of JAVAHERI et al.[12]	2D monitor	×	DSIS
Work of JAVAHERI et al.[27]	2D monitor	×	DSIS
Work of JAVAHERI et al.[15]	2D monitor	×	DSIS
Work of DA SILVA CRUZ et al.[16]	2D monitor	×	DSIS
Work of SU et al.[18]	2D monitor	×	DSIS
IRPC[19]	2D monitor	×	DSIS
WPC[5]	2D monitor	×	DSIS
SJTU-PCQA[6]	2D monitor	×	ACR
VsenseVVDB2[20]	2D monitor	×	ACR
Work of CAO et al.[21]	2D monitor	×	ACR
Work of ALEXIOU and EBRAHIMI[8]	2D monitor	×	DSIS, ACR
VsenseVVDB[17]	2D monitor	×	DSIS, PWC
Work of ZHANG et al.[37]	2D monitor	×	-
Work of ALEXIOU et al.[25]	2D monitor	√	DSIS
Work of ALEXIOU et al.[22]	2D monitor	√	DSIS
LS-PCQA[4]	2D monitor	√	DSIS
Work of TORLIG et al.[10]	2D monitor	√	DSIS
M-PCCD[11]	2D monitor	√	DSIS
Work of ALEXIOU et al.[23]	2D monitor	√	DSIS, ACR
Work of ALEXIOU et al.[24]	2D monitor	√	DSIS, ACR
Work of VIOLA et al.[26]	2D monitor	-	DSIS
NBU-PCD 1.0[28]	2D monitor	-	-
ICIP2020[31]	2D/3D monitor	×	DSIS
RG-PCD[30]	2D/3D monitor	×	DSIS
Work of ALEXIOU et al.[29]	AR	√	DSIS
Work of NEHMÉ et al.[9]	HMD	×	DSIS, ACR
PointXR[35]	HMD	√	DSIS
SIAT-PCQD[36]	HMD	√	DSIS
Work of SUBRAMANYAM et al.[34]	HMD	√	ACR
Work of JESÚS GUTIÉRREZ et al.[38]	HMD	√	ACR

ACR: Absolute Category Rating
AR: augmented reality
DSIS: Double-Stimulus Impairment Scale
HMD: Head-Mounted Display
ICIP2020: A point cloud quality assessment dataset proposed in IEEE International Conference on Image Processing 2020
IRPC: IST (Instituto Superior Téchico) Render Point Cloud Quality Assessment
LS-PCQA: Large Scale Point Cloud Quality Assessment Dataset
M-PCCD: MPEG Point Cloud Compression Dataset
NBU-PCD: Ningbo University Point Cloud Dataset
PointXR: A Point cloud quality assessment dataset developed by PointXR toolbox
PWC: Pairwise Comparison
RG-PCD: Reconstructed Geometry Point Cloud Dataset
SIAT-PCQD: Shenzhen Institute of Advanced Technology Point Cloud Quality Dataset
SJTU-PCQA: Shanghai Jiao Tong University Point Cloud Quality Assessment Dataset Vsense
VVDB: Vsense Volumetric Video Quality Databases
WPC: Waterloo Point Cloud Dataset

ALEXIOU et al. compared the DSIS and ACR methods in Ref. [8]. They found the phenomenon that the DSIS and ACR methods are statistically equivalent, but there are slight differences in the assessment results for different types of distortion. They also found that subjects prefer the DSIS evaluation method. Meanwhile, NEHMÉ et al.[9] extended this study to VR environments, finding that DSIS is more accurate than ACR, especially in terms of color distortion. Their conclusions are highly consistent with the observation in Table 1 that DSIS is more commonly used than ACR and PWC methods. In addition, by observing Table 1, we find that many current subjective evaluations[4 - 6, 8, 10 - 28] are conducted through a 2D monitor, but with the continuous development of display technology and AR/VR technology, emerging display technologies are gradually being applied to subjective experiments. ALEXIOU et al.[29] used augmented reality HMDs for the first time in point cloud quality assessment work. Additionally, 3D monitors were used in the work of ALEXIOU et al.[30 - 31] and ICIP2020 for point cloud subjective evaluation. Although existing studies[32 - 33] have pointed out that 3D display technology is more likely to cause adverse side effects (including dizziness, nausea, disorientation, etc.), we cannot deny that the immersive experience and rich visual information provided by 3D display technology are incomparable to a 2D monitor. Another way to enhance the audience's experience is to introduce interaction. This interaction can freely adjust the viewing angle with a mouse on a 2D monitor[4, 11], or it can freely move around and observe point clouds in an AR/VR environment through HMD devices[34 - 36]. LIU et al.[5] believe that introducing interaction and passive observation of point clouds by the audience are equally effective in subjective tests, but the latter method has a slight advantage in terms of repeatability. In conclusion, the optimal subjective testing setup for point clouds is still an unresolved issue, and in most cases, the appropriate experimental setup is chosen based on the actual situation. Therefore, further exploration in this area is needed.

2.3 Related Subjective Databases

The successive establishment of various databases has played a pivotal role in advancing the field of point cloud quality assessment. As early as the last century, Stanford University established a 3D scanning database[39], which is still used in current PCQA research[8, 29 - 30]. However, the deficiencies of this database have emerged with the gradual deepening of related PCQA research. Firstly, the point cloud content covered in Stanford's 3D scanning library is not diverse enough. As a result, the Motion Picture Experts Group (MPEG) and the Joint Photographic Experts Group (JPEG) proposed the MPEG point cloud database[40] and the JPEG Pleno database[41] respectively, using cultural relics, daily necessities, and human figures as the subjects of point cloud quality research. The point clouds covered in these databases have been continued in many later databases[6, 8, 11, 17, 19 - 20, 28 - 31, 35 - 36, 42]. In the databases established afterward, the content and objects represented by point clouds became more and more abundant, which provides convenience for the development of related research on PCQA.

In recent years, DA SILVA CRUZ et al.[16] and AK et al.[43] have chosen not only the common point cloud representations of humans and animals and inanimate objects during the process of conducting subjective experiments but also included architecture and landscapes. The former introduced eight original point clouds in the experiment to explore the subjective evaluation methods of point clouds, while the latter established the BASICS database[43] around 75 original point clouds and conducted research on objective evaluation methods. Furthermore, because most of the point clouds in the database[39] are colorless, after ALEXIOU et al.[8, 13 - 14, 30] and JAVAHERI et al.[12] established the initial workflow of subjective colorless point cloud evaluation, colored point clouds attracted a wide range of research interest due to their richer visual information, becoming mainstream in PCQA. Since then, various colored point cloud databases have been proposed[4 - 6, 11, 17, 20, 28, 31, 35, 36, 42 - 45] and in-depth research on subjective evaluation has been conducted. ZERMAN et al.[17, 20] established a series of V-SENSE Volumetric Video Quality Databases (vsenseVVDB), and through Unity rendering of volume video of point clouds, they conducted subjective assessments of colored point clouds, finding that texture distortion is more likely to affect point cloud quality than geometric distortion. They also compared the quality of mesh and point cloud, two common types of 3D data representation, during limited bitrate transmission. The results showed that the mesh had higher quality in high bitrate transmission, and point clouds had the opposite conclusion. ALEXIOU et al.[35] also used Unity to develop the PointXR toolbox, but assessed the quality of colored point cloud content through virtual reality technology, enhancing the interactivity of the evaluation process. By analyzing participant interaction patterns, the authors found that, under six degrees-of-freedom (6DoF) observation, participants preferred close-up frontal observation. Real-world applications often require databases to more realistically simulate the possible distortion effects of point clouds in the communication process of collection, storage, compression, transmission, rendering, and display, but the database[39] can no longer meet these requirements.

In the process of establishing later databases, SU et al.[18] applied downsampling, Gaussian noise, and three advanced point cloud compression algorithms to create distorted point clouds. They clarified the types of point cloud distortions from both geometric and texture perspectives. Geometric distortions include hollows, geometric noise, holes, shape distortions, collapses, gaps, and blurring, while texture distortions include texture noise, blocks, blurring, and bleeding. After the MPEG committee standardized the advanced point cloud encoder, the two advanced point cloud codecs, GPCC and VPCC, received more attention in subjective PCQA research. JAVAHERI et al. established the IRPC database, studying the impact of three different encoding methods and three rendering solutions on the visual perceptual quality of point clouds[19]. Additionally, the authors creatively evaluated the performance of the encoding scheme proposed by MPEG. PERRY et al. confirmed that VPCC's compression performance on static content is superior to GPCC through subjective point cloud quality assessment experiments conducted in four independent labs against the established ICIP2020 database[31]. In the databases established in recent years, the types of distortions are more standardized and comprehensive. The WPC database established by LIU et al. covers Gaussian noise, downsampling, GPCC, and VPCC[5]. YANG et al. established the SJTU-PCQA[6], which simulates distortions during the communication process, including octree-based compression, color noise, geometric noise, and scaling enhancement.

To more intuitively and clearly exhibit the development of point cloud subjective databases, we have listed the information of some databases in recent years in Table 2. From Table 2, we can surmise that subjective PCQA databases are striving to develop in the direction of larger scale, more comprehensive distortions, and more realistic and richer point cloud models.

3 Objective Point Cloud Quality Assessment

Although subjective quality evaluation is considered to be the test method that best matches the visual perception of the human eye, conducting subjective experiments often costs a great deal of labor and time. Therefore, objective point cloud quality evaluation has emerged as a promising alternative to alleviate the drawbacks of subjective evaluations. Some objective point cloud quality evaluation methods are listed in Table 3. As with traditional image or video quality assessment, defined from the perspective of the amount of reference information, objective PCQA methods can be categorized into full-reference (FR), reduced-reference (RR), and no-reference (NR) assessment methods. Within this classification, FR PCQA methods are distinguished by their utilization of complete reference point clouds during the assessment of distorted point clouds. Conversely, NR PCQA methods rely exclusively on the distorted point clouds for quality evaluation, without access to the reference point clouds. The RR PCQA methods possess the capability to employ a subset of feature information extracted from the reference point clouds as reference. Subsequently, they conduct a comparative and analytical evaluation of the distorted point clouds, culminating in the derivation of quality assessment outcomes. As defined by the feature extraction method, the objective PCQA methods can

▼Table 2. An overview of subjective PCQA databases

Database	Year	Attribute	Models	Distortion Type
G-PCD[8, 29]	2017	Colorless	40	Octree, Gaussian noise
RG-PCD[30]	2018	Colorless	24	Octree
VsenseVVDB[17]	2019	Colored	32	VPCC
M-PCCD[11]	2019	Colored	244	GPCC, VPCC
IRPC[19]	2020	Colorless & Colored	54 & 54	GPCC, VPCC
ICIP2020[31]	2020	Colored	90	GPCC, VPCC
PointXR[35]	2020	Colored	100	GPCC
NBU-PCD 1.0[28]	2020	Colored	160	Octree
VsenseVVDB2[20]	2020	Colored	164	Draco+JPEG, GPCC, VPCC
SJTU-PCQA[6]	2020	Colored	420	Octree, downsampling, color and geometry noise
SIAT-PCQD[36]	2021	Colored	340	VPCC
CPCD 2.0[42]	2021	Colored	360	GPCC, VPCC, Gaussian noise
WPC[5]	2021	Colored	740	Gaussian noise, downsampling, GPCC, VPCC
WPC2.0[44]	2021	Colored	400	VPCC
WPC3.0[45]	2022	Colored	350	VPCC
LS-PCQA[4]	2023	Colored	1 080	Color and geometry noise, downsampling, GPCC, VPCC, etc.
BASICS[43]	2023	Colored	1 494	VPCC, GPCC, GeoCNN[46]

BASICS: Broad Quality Assessment of Static Point Clouds in Compression Scenarios
CPCD: Color Point Cloud Dataset with GPCC/VPCC Coding and Gaussian Noise Distortions
GeoCNN: Geometry-Based Point Cloud Convolutional Neural Network
GPCC: Geometry-Based Point Cloud Compression
G-PCD: Geometry Point Cloud Database
ICIP2020: A point cloud quality assessment dataset proposed in IEEE International Conference on Image Processing 2020
IRPC: IST (Instituto Superior Téchico) Render Point Cloud Quality Assessment
JPEG: Joint Photographic Experts Group
LS-PCQA: Large Scale Point Cloud Quality Assessment Dataset
M-PCCD: MPEG Point Cloud Compression Dataset
NBU-PCD: Ningbo University Point Cloud Dataset
PointXR: A point cloud quality assessment dataset developed by PointXR toolbox
RG-PCD: Reconstructed Geometry Point Cloud Dataset
SIAT-PCQD: Shenzhen Institute of Advanced Technology Point Cloud Quality Dataset
SJTU-PCQA: Shanghai Jiao Tong University Point Cloud Quality Assessment Dataset
VsenseVVDB: Vsense Volumetric Video Quality Databases
VPCC: Video-based Point Cloud Compression
WPC: Waterloo Point Cloud Dataset

▼Table 3. Summary of objective cloud quality assessment methods

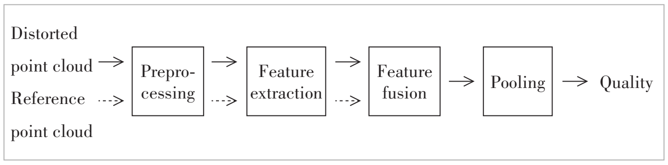
Method	Reference Type	Feature Extraction	Handcrafted/ Deep Learning
p2point[68]	FR	Model-based	Handcrafted
p2plane[47, 49]	FR	Model-based	Handcrafted
p2mesh[69]	FR	Model-based	Handcrafted
Plane to plane[23]	FR	Model-based	Handcrafted
PointSSIM[52]	FR	Model-based	Handcrafted
GraphSIM[55]	FR	Model-based	Handcrafted
MS-GraphSIM[63]	FR	Model-based	Handcrafted
PCQM[53]	FR	Model-based	Handcrafted
PC-MSDM[50]	FR	Model-based	Handcrafted
Proposed by VIOLA et al.[26]	FR	Model-based	Handcrafted
VQA-CPC[28]	FR	Model-based	Handcrafted
CPC-GSCT[42]	FR	Model-based	Handcrafted
Proposed by JAVAHERI et al.[27]	FR	Model-based	Handcrafted
Proposed by JAVAHERI et al.[51]	FR	Model-based	Handcrafted
Proposed by DINIZ et al.[56]	FR	Model-based	Handcrafted
Proposed by DINIZ et al.[57]	FR	Model-based	Handcrafted
Proposed by DINIZ et al.[58]	FR	Model-based	Handcrafted
Proposed by DINIZ et al.[59]	FR	Model-based	Handcrafted
Proposed by DINIZ et al.[60]	FR	Model-based	Handcrafted
EPES[62]	FR	Model-based	Handcrafted
PSNR_{yuv}[10]	FR	Projection-based	Handcrafted
Proposed by WU et al.[36]	FR	Projection-based	Handcrafted
Proposed by HE et al.[65]	FR	Projection-based	Handcrafted
PB-PCQA[6]	FR	Projection-based	Handcrafted
TGP-PCQA[70]	FR	Projection-based	Handcrafted
Proposed by TU et al.[71]	FR	Model & projection	Handcrafted
PCMRR[72]	RR	Model-based	Handcrafted
R-PCQA[73]	RR	Model-based	Handcrafted
RR-CAP[74]	RR	Projection-based	Handcrafted
3D-NSS[64]	NR	Model-based	Handcrafted
StreamPCQ[75]	NR	Model-based	Handcrafted
Proposed by ZHOU et al.[76]	NR	Model-based	Handcrafted
ResSCNN[4]	NR	Model-based	Deep learning
PKT-PCQA[77]	NR	Model-based	Deep learning
Proposed by TU et al.[78]	NR	Projection-based	Deep learning
GPA-Net[79]	NR	Projection-based	Deep learning
PQA-Net[67]	NR	Projection-based	Deep learning
GMS-3DQA[80]	NR	Projection-based	Deep learning
D^3-PCQA[81]	NR	Projection-based	Deep learning
PM-BVQA[66]	NR	Projection-based	Deep learning
IT-PCQA[82]	NR	Projection-based	Deep learning
3D-CNN-PCQA[83]	NR	Projection-based	Deep learning
VQA-PC[84]	NR	Projection-based	Deep learning
BQE-CVP[61]	NR	Model & projection	Handcrafted
MM-PCQA[85]	NR	Model & projection	Deep learning

CAP: content-oriented saliency projection
3D-CNN-PCQA: 3 Dimension Convolutional Neural Network Point Cloud Quality Assessment
3D-NSS: 3 Dimension Natural Scene Statistics
BQE-CVP: Blind quality evaluator for colored point cloud based on visual perception
CPC-GSCT: A quality assessment metric for colored point cloud based on geometric segmentation and color transformation
D3-PCQA: Point cloud quality assessment via domain-relevance degradation description
EPES: Point cloud quality modeling using elastic potential energy similarity
FR: full-reference assessment
GMS-3DQA: Projection-based grid mini-path sampling for 3D model quality assessment
GPA: Graph Convolutional Point Cloud Assessment
GraphSIM: Graph Similarity
IT-PCQA: Image Transferred Point Cloud Quality Assessment
MM-PCQA: Multi-Modal Point Cloud Quality Assessment
MS-GraphSIM: Multi-Scale Graph Similarity
NR: no-reference assessment
p2mesh: Point to Mesh
p2plane: Point to Plane
p2point: Point to Point
PB-PCQA: Projection-Based Point Cloud Quality Assessment
PCMRR: A reduced reference metric for visual quality evaluation of point cloud content
PC-MSDM: Point Cloud-Mesh Structural Distortion Measure
PCQM: Point Cloud Quality Metric
PKT-PCQA: Progressive knowledge transfer based on human visual perception mechanism for point cloud quality assessment
PM-BVQA: Point cloud projection and multi-scale feature fusion network based blind visual quality assessment
PointSSIM: Point Cloud Structure Similarity Index Measure
PQA: Point Cloud Quality
PSNRyuv: Peak Signal-to-Noise Ratio in Yuv
ResSCNN: Residual Sparse Convolutional Neural Network
R-PCQA: Reduced Reference Point Cloud Quality Assessment
RR: reduced-reference assessment
PCQ: an overall bitstream-based point cloud quality assessment
VQA-CPC: Visual quality assessment metric of color point clouds
VQA-PC: Dealing with point cloud quality assessment tasks via using video quality assessment

be further categorized into two main groups: model-based methods[4, 23, 26 – 28, 42, 47 – 64] and projection-based methods[6, 10 – 11, 18, 25, 36, 65 – 67].

3.1 Model-Based Methods

In the early stage of research on objective point cloud quality assessment, most of the research started from the perspective of the 3D model of the point cloud. Fig. 3 illustrates the general framework of the model-based methods. Specifically, methods such as p2point[68], p2plane[47, 49] p2mesh[69] and plane2plane[23] give quality scores by calculating the distance between discrete points as a similarity metric. Differently, p2point[68] uses the Euclidean distance between points as a similarity measure, p2plane[47, 49] calculates the projection error of related points along the discovery direction, while plane2plane[23] evaluates the quality of point clouds through the angle similarity of the tangent plane. Later, JAVAHERIE et al. introduced more distance measures into objective PCQA, including Peak-Signal-to-Noise Ratio (PSNR)[27], Generalized Hausdorff Distance[51] and Mahalanobis Distance[54], to effectively measure the correspondence between points and distributions. MEYNET et al.[50] extended the Mesh Structural Distortion Measure (MSDM)[86 – 87] metric method in Mesh to the point cloud field, and designed PC-MSDM based on local curvature statistics. However, these works still stay in the measurement of geometric features of point clouds and ignore the



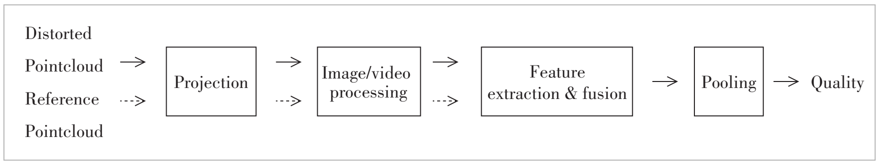
▲Figure 3. General framework of model-based point cloud quality assessment (PCQA) methods. Dashed lines indicate different amounts of reference information in full-reference (FR), reduced-reference (RR), and no-reference (NR) methods rich color information that point clouds have.

In the PCQA standards collected by MPEG, in addition to p2point[68] and p2plane[47, 49], there is also PSNR$_{yuv}$[10] that can perceive the texture distortion of colored point clouds. Since this method is based on the expansion of PSNR, it inevitably possesses the limitations of PSNR itself[88 - 89]. Therefore, ALEXIOU et al.[52] considered extending Structural Similarity (SSIM)[88]. They focused on studying four factors: geometry, normal vector, curvature, and color, and proposed PointSSIM. Meanwhile, MEYNET et al.[53] explored four factors in Point Cloud Quality Metric (PCQM), namely, curvature, brightness, chroma, and hue. These factors are combined by using optimal linear weighting. Furthermore, DINIZ et al. further explored point cloud color perception and proposed statistical variants of local binary patterns (LBP)[56 - 57], perceived color distance patterns (PCDP)[58], and local luminance patterns (LLP)[59], achieving excellent performance on multiple databases. Besides, they combined geometric and color statistical information to propose a low-complexity BitDance[60] algorithm. These methods combine the geometric and color information of point clouds and promote the development of objective quality assessment of colored point clouds to a certain extent. However, whether the underlying principles of these methods are in line with the human visual system still needs to be further verified. YANG et al.[55] proposed GraphSIM based on the fact that the human visual system is more sensitive to high-frequency signals, integrating local graph similarity features and color gradients. ZHANG et al.[63] improved GraphSIM with consideration to the multi-scale characteristic of human visual perception and proposed MS-GraphSIM. We can see that many current works are paying more attention to the perception mechanism of the human visual system itself. On the other hand, we also see XU et al.[62] combining the concept of elastic potential energy similarity, interpreting point cloud distortion as the work done by external forces on the reference point cloud, and creatively combining relevant knowledge in the physical field with visual perception. Whether the rich research results in other fields can guide objective point cloud quality assessment is a topic worthy of attention and in-depth exploration.

The above methods all involve reference point clouds when evaluating point cloud quality, and in many practical cases, we cannot obtain all the reference point clouds or there are no point clouds for reference at all. Therefore, related research on RR and NR assessment methods is very necessary. Due to the lack of reference information, RR and NR objective PCQA methods are more challenging. A Reduced Reference Metric for Visual Quality Evaluation of Point Cloud Content (PCMRR)[72] and Reduced reference Point Cloud Quality Assessment (R-PCQA)[73] are two commonly used model-based RR PCQA methods. The former reduces references by extracting statistical features in the geometry, color and normal vector domains, while the latter achieves reduced-reference by fitting the relationship between quantization steps and perceived quality. Feature extraction can be divided into two types. One is to manually extract the required features based on the model itself. For example, ZHOU et al.[76] combined human brain cognition to design a blind evaluation method using a structure-guided resampling (SGR) method, extracting three features: ensemble density, color naturalness, and angle consistency. ZHANG et al.[64] used 3D scene statistics (3D-NSS) and entropy to extract quality perception features, and finally used support vector regression (SVR) to get the quality score of the point cloud. SU et al.[75] explored from the perspective of end-user Quality of Experience (QoE) and developed a bitstream-based no-reference method. Another type is to use deep learning to extract point cloud quality features. Typical methods include ResSCNN[4] and perceptual quality assessment of point clouds (PKT-PCQA)[77]. The former is based on sparse convolutional neural networks and the latter is based on progressive knowledge transfer. Combined with Table 3, it is not difficult to find that existing works rarely extract quality features of point cloud models themselves through deep learning. One possible reason is that point clouds, as a dense data structure, require a huge amount of space and cost in storage and calculation. Therefore, point clouds are not suitable for direct feature extraction through deep learning.

3.2 Projection-Based Methods

As shown in Fig. 4, the projection-based method projects a 3D point cloud and represents the quality of the entire point cloud by evaluating the quality of the projection. The method effectively circumvents the problems of storage space and computational overhead caused by the point cloud. Projection-based methods can be used for full-reference objective point cloud quality evaluation[6, 36, 65], as well as reduced-reference[74] and no-reference quality evaluation methods[66 - 67, 78 - 84]. Regarding the setup of projection, ALEXIOU et al. conducted experiments in Ref. [25]. The results show that when the projection exceeds six projection planes, the quality prediction performance does not significantly improve. Based on these results, YANG et al.[6] first projected the reference point cloud and distorted point clouds onto six planes separately through perspective projection, and then extracted global and local features of depth and color images through projection to evaluate



▲ **Figure 4. General framework of projection-based point cloud quality assessment (PCQA) methods. Dashed lines indicate different amounts of reference information in full-reference (FR), reduced-reference (RR), and no-reference (NR) methods** point cloud quality. However, WU et al. [36] believe that this method causes inevitable occlusion and misalignment in the point cloud during the projection process. In addition, they believe that projections from different angles have different impacts on visual perception. Therefore, they proposed a view-based projection weighted model and a block-based projection model. ZHOU et al. [74] applied the projection method to reduced-reference point cloud quality evaluation. They simplified the reference point cloud and distorted point cloud through downsampling to obtain content-oriented saliency projection (RR-CAP), so that users do not need to obtain a large number of reference point clouds from the transmission end when evaluating point cloud quality. In contrast to model-based objective quality evaluation, many no-reference quality evaluation methods based on projection extract features of the projection through deep learning. This is partly because the projection method converts three-dimensional point clouds into two-dimensional data for processing, reducing the computational complexity and making deep learning feasible in point cloud quality evaluation. On the other hand, due to the excellent performance of deep learning in many computer vision tasks, scholars unanimously acknowledge the outstanding feature extraction capability of deep networks. The specific implementation methods include evaluating point clouds by projecting point clouds into images and using existing image quality evaluation methods[83 – 84, 90 – 97], and processing point clouds rendered into videos from different angles by setting the orbit of virtual cameras. The former extracts temporal features from rendered point cloud videos using a modified ResNet3D[98], while the latter believes that temporal features are insufficient to describe the quality of point clouds, so it selects key frames from point cloud videos for spatial feature extraction using 2D-CNN and finally evaluates point clouds combining temporal and spatial features. In addition, hot topics in the field of deep learning such as multimodal learning[66], multitask learning[67, 79], dual-stream convolutional networks[78], graph convolutional networks[79], domain adaptation[82], and domain generalization[81] have also been applied in no-reference point cloud quality evaluation. By observing Table 3, we can find that Refs. [61] and [85] combine the two mainstream methods of model-based and projection-based to evaluate the quality of point clouds. After analysis, we can conclude that although projection-based methods have significant advantages in efficiency and computational quantity, they inherently observe three-dimensional point clouds through two-dimensional virtual cameras, inevitably leading to the problem of incomplete information observation. Therefore, to alleviate the limitations of two-dimensional media, it is feasible and worth exploring to introduce model-based methods. However, at the same time, how to effectively weigh the computational overhead and evaluate the performance of the method is also a topic that needs to be addressed and explored in depth.

4 Evaluation of PCQA Models

4.1 Evaluation Protocol

The current point cloud quality evaluation methods generally follow the recommendations given by the Video Quality Expert Group (VQEG)[99] in the field of image quality assessment (IQA). The evaluation is conducted from three aspects: prediction accuracy, monotonicity, and consistency. Typically, a five-parameter monotonic logistic function is used to calculate the quality score:

$$p = \beta_1\left(0.5 - \frac{1}{1 + e^{\beta_2(o - \beta_3)}}\right) + \beta_4 o + \beta_5, \tag{1}$$

where o and p represent the predicted scores and mapped scores, respectively. After nonlinear mapping, the performance of the PCQA model can be measured using the following four commonly used criteria: Spearman Rank-order Correlation Coefficient (SRCC), Pearson Linear Correlation Coefficient (PLCC), Kendall Rank-order Correlation Coefficient (KRCC), and Root Mean Square Error (RMSE). Eq. 2 provides the calculation process of SRCC:

$$\mathrm{SRCC} = 1 - \frac{6\sum_{n=1}^{N} d_i^2}{N\left(N^2 - 1\right)}, \tag{2}$$

where d_i represents the difference in rankings between the objective score and predicted score of the i-th point cloud, and N represents the total number of point clouds. SRCC is used to measure the monotonicity of visual quality prediction, with its value ranging from 0 to 1. The closer SRCC is to 1, the better the performance of the model is considered to be. Eq. 3 provides the calculation process of PLCC:

$$\mathrm{PLCC} = \frac{\sum_{i=1}^{N}\left(p_i - \bar{p}\right)\left(s_i - \bar{s}\right)}{\sqrt{\sum_{i=1}^{N}\left(p_i - \bar{p}\right)^2\left(s_i - \bar{s}\right)^2}}, \tag{3}$$

where s_i and p_i indicate the objective score and predicted score of the i-th point cloud, and $\bar{s}$ and $\bar{p}$ stand for the average values for s_i and p_i. PLCC is used to measure the linearity and consistency of visual quality prediction results, with its value ranging from 0 to 1. The closer PLCC is to 1, the better the performance of the model is considered to be. Eq. 4 provides the calculation process of KRCC:

$$\mathrm{KRCC} = \frac{N_c - N_d}{0.5(N - 1)N}, \tag{4}$$

where N_c and N_d represent the number of consistent pairs and discordant pairs. KRCC utilizes the concept of “paired” data to determine the strength of the correlation coefficient. It can also be used to describe the monotonicity of visual quality prediction, with its value ranging from 0 to 1. The closer KRCC is to 1, the better the performance of the model is considered to be. Eq. 5 provides the calculation process of RMSE:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(p_i - s_i\right)^2}, \tag{5}$$

where s_i and p_i similarly represent the subjective score and the objective score after nonlinear mapping of the i-th point cloud. RMSE can be used to measure the accuracy of visual quality prediction. The lower the RMSE value, the better the performance of the model is considered to be.

4.2 Performance Comparison

In this section, we summarize the performance of common PCQA methods. Since not all surveyed methods are publicly available, we did not include all of them in the evaluation. We selected four widely used PCQA databases: IRPC[19], CPCD2.0[42], SJTU-PCQA[6], and WPC[5], to test all participating evaluation methods. Specifically, we reported the SRCC, PLCC, KRCC and RMSE metrics for full-reference, reduced-reference, and no-reference quality assessment methods in Tables 4 and 5. It is worth stating that this chapter also reports some commonly used image and video quality evaluation methods to provide a more comprehensive assessment of objective point cloud quality evaluation methods. This is because projection-based methods are essentially a dimensionality reduction process, allowing 3D point clouds to be evaluated by 2D image or video quality evaluation methods as well.

Combining Tables 4 and 5, we can see that most of the methods with optimal performance take into account the features provided by the 3D model itself. This result proves that 3D models do provide more effective quality features than 2D projections. On the other hand, we can also see that the current projection-based methods represented by VQA-PC[84] have also achieved notable performance. One possible explanation is that VQA-PC focuses on dynamic quality-aware information, and recording the point cloud as a video through four moving paths allows for more point cloud detail from different viewpoints, thus preserving as much of the point

▼Table 4. Performance comparison of different PCQA methods on IRPC and CPCD2.0. For FR and NR methods, the best performance of each metric is marked in bold and underlined bold respectively. The IQA and VQA methods are marked with * superscript

Reference	Type	Methods	IRPC				CPCD2.0			
			SRCC	PLCC	KRCC	RMSE	SRCC	PLCC	KRCC	RMSE
FR	Model-based	p2point$_{\text{Hausdorff}}$[68]	0.212 5	0.238 8	0.145 5	0.960 1	0.314 5	0.348 2	0.217 9	1.099 5
		p2point$_{\text{MSE}}$[68]	0.328 1	0.335 7	0.214 6	0.931 3	0.549 1	0.678 4	0.414 2	0.861 7
		p2plane$_{\text{Hausdorff}}$[47, 49]	0.254 1	0.392 5	0.197 5	0.908 9	0.378 6	0.406 1	0.266 3	1.071 8
		p2plane$_{\text{MSE}}$[47, 49]	0.256 4	0.429 6	0.195 7	0.892 8	0.569 2	0.691 4	0.438 5	0.847 4
		AS$_{\text{MEAN}}$[23]	0.112 3	0.156 9	0.066 9	0.976 4	0.404 4	0.437 6	0.275 2	1.054 6
		AS$_{\text{RMS}}$[23]	0.118 8	0.145 2	0.085 2	0.978 2	0.417 3	0.446 4	0.289 5	1.049 6
		AS$_{\text{MSE}}$[23]	0.118 8	0.153 6	0.085 2	0.990 2	0.417 3	0.447 2	0.289 5	1.049 1
		PC-MSDM[50]	0.151 9	0.272 9	0.106 3	0.951 5	0.532 1	0.625 4	0.384 2	0.915 2
		PCQM[53]	0381 9	0.561 1	0.303 3	0.818 4	0.340 8	0.481 3	0.261 5	1.028 1
		CPC-GSCT[42]	**0.862 6**	**0.870 6**	**0.689 4**	**0.482 9**	0.906 3	0.904 9	0.745 1	0.502 7
	Projection-based	PSNR*	0.149 6	0.347 1	0.089 4	0.927 2	0.406 4	0.418 3	0.286 7	1.065 4
		SSIM*[88]	0.080 6	0.238 5	0.048 6	0.960 1	0.534 7	0.564 7	0.379 2	0.968 0
		MS-SSIM*[100]	0.116 4	0.328 0	0.069 7	0.934 0	0.568 6	0.621 2	0.414 0	0.919 2
		VIF*[101]	0.171 6	0.094 9	0.121 7	0.984 2	0.674 4	0.698 5	0.495 7	0.839 4
		TGP-PCQA[70]	0.650 0	0.800 5	0.555 6	0.491 4	**0.906 6**	**0.909 4**	**0.758 9**	**0.489 2**
NR	Model-based & Projection-based	BQE-CVP[61]	**<u>0.729 8</u>**	**<u>0.726 5</u>**	**<u>0.542 7</u>**	**<u>0.658 6</u>**	**<u>0.789 0</u>**	**<u>0.795 0</u>**	**<u>0.598 3</u>**	**<u>0.721 8</u>**

AS: Angular Similarity
BQE-CVP: Blind Quality Evaluator for Colored Point Cloud based on Visual Perception
CPCD: Color Point Cloud Dataset with GPCC/VPCC Coding and Gaussian Noise Distortions
CPC-GSCT: A quality assessment metric for colored point cloud based on geometric segmentation and color transformation
FR: Full Reference
IQA: Image Quality Assessment
IRPC: IST (Instituto Superior Téchico) Render Point Cloud Quality Assessment
KRCC: Kendall Rank-order Correlation Coefficient MSE: Mean Square Error
MS-SSIM: Multi-Scale Structure Similarity Index Measure
NR: No Reference
P2plane: Point to Plane
P2point: Point to Point
PC-MSDM: Point Cloud-Mesh Structural Distortion Measure
PCQA: Point Cloud Quality Assessment
PCQM: Point Cloud Quality Metric
PLCC: Pearson Linear Correlation Coefficient
PSNR: Peak Signal-to-Noise Ratio
RMS: Root Mean Squared
RMSE: Root Mean Square Error
SRCC: Spearman Rank-order Correlation Coefficient
SSIM: Structure Similarity Index Measure
TGP-PCQA: Texture and Geometry Projection based Point Cloud Quality Assessment
VIF: Visual Information Fidelity
VQA: Video Quality Assessment

▼Table 5. Performance comparison of different PCQA methods on SJTU-PCQA and WPC. For FR, RR, and NR methods, the best performance of each metric is marked in bold, bold italics, and underlined bold (vacant metrics are not counted in the comparison) respectively. The IQA and VQA methods are marked with * superscript

Reference	Type	Method	SJTU-PCQA				WPC			
			SRCC	PLCC	KRCC	RMSE	SRCC	PLCC	KRCC	RMSE
FR	Model-based	p2point$_{Hausdorff}$[68]	0.43	0.16	0.30	2.39	0.27	0.39	0.19	20.89
		p2point$_{MSE}$[68]	0.40	0.47	0.28	2.13	0.45	0.48	0.31	19.89
		p2plane$_{Hausdorff}$[47, 49]	0.46	0.37	0.33	2.44	0.28	0.27	0.16	21.98
		p2plane$_{MSE}$[47, 49]	0.49	0.56	0.35	2.00	0.32	0.26	0.22	22.82
		AS$_{MEAN}$[23]	0.51	0.65	0.36	1.82	-	-	-	-
		AS$_{RMS}$[23]	0.52	0.65	0.37	1.82	-	-	-	-
		AS$_{MSE}$[23]	0.52	0.65	0.37	1.82	-	-	-	-
		PC-MSDM[50]	0.32	0.41	0.21	2.21	-	-	-	-
		PCQM[53]	0.74	0.77	0.56	1.52	**0.74**	**0.74**	**0.56**	**15.16**
		GraphSIM[55]	0.84	0.84	0.64	1.57	0.58	0.61	0.41	17.19
		PointSSIM[52]	0.68	0.71	0.49	1.70	0.45	0.46	0.32	20.27
		CPC-GSCT[42]	**0.89**	**0.91**	**0.71**	**0.99**	-	-	-	-
	Projection-based	PSNR$_{yuv}$[10]	-	-	-	-	0.44	0.53	0.31	19.31
		PSNR*	0.65	0.63	0.47	1.87	0.42	0.48	0.30	15.81
		SSIM*[88]	0.55	0.56	0.39	1.99	0.38	0.49	0.32	15.77
		MS-SSIM*[100]	0.72	0.74	0.52	1.62	-	-	-	-
		VIF*[101]	0.74	0.78	0.54	1.49	-	-	-	-
		PB-PCQA[6]	0.60	0.60	-	1.86	-	-	-	-
		TGP-PCQA[70]	0.83	0.86	0.65	1.21	-	-	-	-
RR	Model-based	R-PCQA[73]	-	-	-	-	**0.88**	**0.88**	-	-
		PCMRR[72]	0.48	0.61	0.33	1.93	0.30	0.34	0.20	21.53
	Projection-based	RR-CAP[74]	**0.75**	**0.76**	**0.55**	**1.55**	0.71	0.73	**0.52**	**15.64**
NR	Model-based	3D-NSS[64]	0.71	0.73	0.51	1.76	0.64	0.65	0.44	16.57
	Projection-based	BRISQUE*[91]	0.20	0.22	0.11	2.24	0.37	0.41	0.24	22.54
		NIQE*[96]	0.22	0.37	0.15	2.26	0.38	0.39	0.25	22.55
		IL-NIQE*[93]	0.08	0.16	0.05	2.33	0.09	0.14	0.08	24.01
		VIIDEO*[102]	0.05	0.29	0.04	2.31	0.07	0.08	0.05	22.92
		V-BLIINDS*[103]	0.68	0.78	0.48	1.50	0.46	0.49	0.30	19.73
		TLVQM*[104]	0.52	0.60	0.34	1.91	0.03	0.01	0.20	22.14
		VIDEVAL*[105]	0.60	0.74	0.42	1.50	0.37	0.26	0.36	21.09
		VSFA*[106]	0.72	0.82	0.54	1.40	0.63	0.63	0.46	17.23
		RAPIQUE*[107]	0.44	0.40	0.34	2.21	0.27	0.35	0.20	21.14
		StairVQA*[108]	0.79	0.78	0.55	1.42	0.72	0.71	0.52	15.07
		PQA-Net[67]	-	-	-	-	0.69	0.70	0.51	15.18
		3D-CNN-PC-QA[83]	0.83	0.86	0.60	1.22	0.75	0.76	0.56	13.56
		ResSCNN[4]	0.81	0.86	-	-	-	-	-	-
		IT-PCQA[82]	0.63	0.58	-	-	0.54	0.55		
		VQA-PC[84]	0.85	0.86	0.65	1.13	0.79	0.79	0.61	13.62
	Model-based & projection-based	BQE-CVP[61]	0.89	0.91	0.73	0.97	-	-	-	-
		MM-PCQA[85]	**0.91**	**0.92**	**0.78**	**0.77**	**0.83**	**0.83**	**0.64**	**12.84**

CAP: content-oriented saliency projection
3D-CNN-PCQA: 3 Dimension Convolutional Neural Network Point Cloud Quality Assessment
3D-NSS: 3 Dimension Natural Scene Statistics
AS: Angular Similarity
BQE-CVP: Blind Quality Evaluator for Colored Point Cloud Based on Visual Perception
BRISQUE: Blind/Referenceless Image Spatial Quality Evaluator
CPC-GSCT: A quality assessment metric for colored point cloud based on geometric segmentation and color transformation
FR: full-reference assessment
GraphSIM: Graph Similarity
IL-NIQE: Integrated Local Natural Image Quality Evaluator
IQA: Image Quality Assessment
IT-PCQA: Image Transferred Point Cloud Quality Assessment
KRCC: Kendall Rank-order Correlation Coefficient
MM-PCQA: Multi-Modal Point Cloud Quality Assessment
MSE: Mean Square Error
MS-SSIM: Multi-Scale Structure Similarity Index Measure
NIQE: Natural Image Quality Evaluator
NR: no-reference assessment
p2plane: Point to Plane
p2point: Point to Point
PB-PCQA: Projection-Based Point Cloud Quality Assessment
PCMRR: A Reduced Reference Metric for Visual Quality Evaluation of Point Cloud Content
PC-MSDM: Point Cloud-Mesh Structural Distortion Measure
PCQM: Point Cloud Quality Metric
PLCC: Pearson Linear Correlation Coefficient
PointSSIM: Point Cloud Structure Similarity Index Measure
PQA: Point Cloud Quality
PSNR: Peak Signal-to-Noise Ratio
RAPIQUE: Rapid and accurate video quality prediction of user generated content
ResSCNN: Residual Sparse Convolutional Neural Network
RMS: Root Mean Squared
RMSE: Root Mean Square Error
R-PCQA: Reduced reference Point Cloud Quality Assessment
RR: reduced-reference assessment
SJTU-PCQA: Shanghai Jiao Tong University Point Cloud Quality Assessment Dataset
SRCC: Spearman Rank-order Correlation Coefficient
SSIM: Structure Similarity Index Measure
StairVQA: Staircase Video Quality Assessment
TGP-PCQA: Texture and Geometry Projection Based Point Cloud Quality Assessment
TLVQM: Two-Level Approach for Consumer Video Quality assessment
V-BLIINDS: Blind prediction of natural video quality
VIDEVAL: Video Quality Evaluator
VIF: Visual Information Fidelity
VIIDEO: Video Intrinsic Integrity and Distortion Evaluation Oracle
VQA: Video Quality Assessment
VQA-PC: Dealing with point cloud quality assessment tasks via using video quality assessment
VSFA: a method for quality assessment of in-the-wild videos
WPC: Waterloo Point Cloud Dataset

cloud’s 3D features as possible. Moreover, among the FR methods, we notice that CPC-GSCT[42] performs well on three datasets. We believe that CPC-GSCT can perceive the quality of point clouds in a more comprehensive way by taking into account the geometric properties and color features of point clouds from the perspective of the point cloud model. Besides, among the NR methods, MM-PCQA[85] stands out in terms of performance, which is a novel multimodal fusion method for PCQA. The excellent performance further demonstrates the feasibility of multimodality in PCQA.

5 Conclusions

In this survey, we present a comprehensive and up-to-date review of PCQA. The paper begins with an introduction to point clouds and their wide range of applications. Along with it, there is an increasing demand for point cloud quality. Therefore, point cloud quality assessment has become a topic of great interest. Then we review the subjective quality evaluation of point clouds from three aspects: common distortion types of point clouds, commonly used subjective experimental setups, and existing subjective datasets. However, conducting subjective experiments is costly. Therefore, we further discusses objective point cloud quality evaluation methods, including model-based and projection-based methods. To assess these objective methods, we provide the evaluation criteria and report the performance of multiple approaches on four datasets. Overall, point cloud quality evaluation requires further research and exploration by relevant researchers and practitioners in both subjective and objective methods.

References

[1] CAO C, PREDA M, ZAHARIA T. 3D point cloud compression: a survey [C]//24th International Conference on 3D Web Technology. ACM, 2019: 1 – 9. DOI: 10.1145/3329714.3338130

[2] CHARLES R Q, HAO S, MO K C, et al. PointNet: deep learning on point sets for 3D classification and segmentation [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017: 77 – 85. DOI: 10.1109/CVPR.2017.16

[3] GUO Y L, WANG H Y, HU Q Y, et al. Deep learning for 3D point clouds: a survey [J]. IEEE transactions on pattern analysis and machine intelligence, 2021, 43(12): 4338 – 4364. DOI: 10.1109/TPAMI.2020.3005434

[4] LIU Y P, YANG Q, XU Y L, et al. Point cloud quality assessment: dataset construction and learning-based no-reference metric [J]. ACM transactions on multimedia computing, communications, and applications, 2023, 19(2s): No.80. DOI: 10.1145/3550274

[5] LIU Q, SU H L, DUANMU Z F, et al. Perceptual quality assessment of colored 3D point clouds [J]. IEEE transactions on visualization and computer graphics, 2023, 29(8): 3642 – 3655. DOI: 10.1109/TVCG.2022.3167151

[6] YANG Q, CHEN H, MA Z, et al. Predicting the perceptual quality of point cloud: a 3D-to-2D projection-based exploration [J]. IEEE transactions on multimedia, 2021, 23: 3877 – 3891. DOI: 10.1109/TMM.2020.3033117

[7] LAZZAROTTO D, TESTOLINA M, EBRAHIMI T. On the impact of spatial rendering on point cloud subjective visual quality assessment [C]//14th International Conference on Quality of Multimedia Experience (QoMEX). IEEE, 2022: 1 – 6. DOI: 10.1109/QoMEX55416.2022.9900898

[8] ALEXIOU E, EBRAHIMI T. On the performance of metrics to predict quality in point cloud representations [C]//Proc. SPIE 10396, Applications of Digital Image Processing XL. SPIE, 2017: 282 – 297. DOI: 10.1117/12.2275142

[9] NEHMÉ Y, FARRUGIA J P, DUPONT F, et al. Comparison of subjective methods, with and without explicit reference, for quality assessment of 3D graphics [C]//ACM Symposium on Applied Perception. ACM, 2019: 1 – 9. DOI: 10.1145/3343036.3352493

[10] EBRAHIMI T, ALEXIOU E, FONSECA T A, et al. A novel methodology for quality assessment of voxelized point clouds [C]//Proc. Applications of Digital Image Processing XLI. SPIE, 2018. DOI: 10.1117/12.2322741

[11] ALEXIOU E, VIOLA I, BORGES T M, et al. A comprehensive study of the rate-distortion performance in MPEG point cloud compression [J]. APSIPA transactions on signal and information processing, 2019, 8(1): e27. DOI: 10.1017/atsip.2019.20

[12] JAVAHERI A, BRITES C, PEREIRA F, et al. Subjective and objective quality evaluation of 3D point cloud denoising algorithms [C]//2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). IEEE, 2017: 1 – 6. DOI: 10.1109/ICMEW.2017.8026263

[13] ALEXIOU E, EBRAHIMI T. Impact of visualisation strategy for subjective quality assessment of point clouds [C]//2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). IEEE, 2018: 1 – 6. DOI: 10.1109/ICMEW.2018.8551498

[14] ALEXIOU E, PINHEIRO A M G, DUARTE C, et al. Point cloud subjective evaluation methodology based on reconstructed surfaces [C]//Proc. SPIE 10752, Applications of Digital Image Processing XLI. SPIE, 2018, 10752: 160 – 173. DOI: 10.1117/12.2321518

[15] JAVAHERI A, BRITES C, PEREIRA F, et al. Subjective and objective quality evaluation of compressed point clouds [C]//IEEE 19th International Workshop on Multimedia Signal Processing (MMSP). IEEE, 2017: 1 – 6. DOI: 10.1109/MMSP.2017.8122239

[16] DA SILVA CRUZ L A, DUMIĆ E, ALEXIOU E, et al. Point cloud quality evaluation: towards a definition for test conditions [C]//2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX). IEEE, 2019: 1 – 6. DOI: 10.1109/QoMEX.2019.8743258

[17] ZERMAN E, GAO P, OZCINAR C, et al. Subjective and objective quality assessment for volumetric video compression [J]. Electronic imaging, 2019, 31 (10): 323 – 1. DOI: 10.2352/issn.2470-1173.2019.10.iqsp-323

[18] SU H L, DUANMU Z F, LIU W T, et al. Perceptual quality assessment of 3d point clouds [C]//2019 IEEE International Conference on Image Processing (ICIP). IEEE, 2019: 3182 – 3186. DOI: 10.1109/ICIP.2019.8803298

[19] JAVAHERI A, BRITES C, PEREIRA F, et al. Point cloud rendering after coding: impacts on subjective and objective quality [J]. IEEE transactions on multimedia, 2021, 23: 4049 – 4064. DOI: 10.1109/TMM.2020.3037481

[20] ZERMAN E, OZCINAR C, GAO P, et al. Textured mesh vs coloured point cloud: a subjective study for volumetric video compression [C]//Twelfth International Conference on Quality of Multimedia Experience (QoMEX). IEEE, 2020: 1 – 6. DOI: 10.1109/QoMEX48832.2020.9123137

[21] CAO K M, XU Y, COSMAN P. Visual quality of compressed mesh and point cloud sequences [J]. IEEE access, 2020, 8: 171203 – 171217. DOI: 10.1109/ACCESS.2020.3024633

[22] ALEXIOU E, EBRAHIMI T. On subjective and objective quality evaluation of point cloud geometry [C]//Ninth International Conference on Quality of Multimedia Experience (QoMEX). IEEE, 2017: 1 – 3. DOI: 10.1109/QoMEX.2017.7965681

[23] ALEXIOU E, EBRAHIMI T. Point cloud quality assessment metric based on angular similarity [C]//2018 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2018: 1 – 6. DOI: 10.1109/ICME.2018.8486512

[24] ALEXIOU E, EBRAHIMI T. Benchmarking of objective quality metrics for colorless point clouds [C]//2018 Picture Coding Symposium (PCS). IEEE, 2018: 51 – 55. DOI: 10.1109/PCS.2018.8456252

[25] ALEXIOU E, EBRAHIMI T. Exploiting user interactivity in quality assessment of point cloud imaging [C]//Eleventh International Conference on Quality of Multimedia Experience (QoMEX). IEEE, 2019: 1 – 6. DOI: 10.1109/QoMEX.2019.8743277

[26] VIOLA I, SUBRAMANYAM S, CESAR P. A color-based objective quality metric for point cloud contents [C]//Twelfth International Conference on Quality of Multimedia Experience (QoMEX). IEEE, 2020: 1 – 6. DOI: 10.1109/QoMEX48832.2020.9123089

[27] JAVAHERI A, BRITES C, PEREIRA F, et al. Improving PSNR-based quality metrics performance for point cloud geometry [C]//2020 IEEE International Conference on Image Processing (ICIP). IEEE, 2020: 3438 – 3442. DOI: 10.1109/ICIP40778.2020.9191233

[28] HUA L, YU M, JIANG G Y, et al. VQA-CPC: a novel visual quality assessment metric of color point clouds [C]//Proc. SPIE 11550, Optoelectronic Imaging and Multimedia Technology VII. SPIE, 2020, 11550: 244 – 252. DOI: 10.1117/12.2573686

[29] ALEXIOU E, UPENIK E, EBRAHIMI T. Towards subjective quality assessment of point cloud imaging in augmented reality [C]//IEEE 19th International Workshop on Multimedia Signal Processing (MMSP). IEEE, 2017: 1 – 6. DOI: 10.1109/MMSP.2017.8122237

[30] ALEXIOU E, EBRAHIMI T, BERNARDO M V, et al. Point cloud subjective evaluation methodology based on 2D rendering [C]//Tenth International Conference on Quality of Multimedia Experience (QoMEX). IEEE, 2018: 1 – 6. DOI: 10.1109/QoMEX.2018.8463406

[31] PERRY S, CONG H P, DA SILVA CRUZ L A, et al. Quality evaluation of static point clouds encoded using MPEG codecs [C]//2020 IEEE International Conference on Image Processing (ICIP). IEEE, 2020: 3428 – 3432. DOI: 10.1109/ICIP40778.2020.9191308

[32] SOLIMINI A G. Are there side effects to watching 3D movies? A prospective crossover observational study on visually induced motion sickness [J]. PLoS one, 2013, 8(2): e56160. DOI: 10.1371/journal.pone.0056160

[33] SHARPLES S, COBB S, MOODY A, et al. Virtual reality induced symptoms and effects (VRISE): comparison of head mounted display (HMD), desktop and projection display systems [J]. Displays, 2008, 29(2): 58 – 69. DOI: 10.1016/j.displa.2007.09.005

[34] SUBRAMANYAM S, LI J, VIOLA I, et al. Comparing the quality of highly realistic digital humans in 3DoF and 6DoF: a volumetric video case study [C]// 2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR). IEEE, 2020: 127 – 136. DOI: 10.1109/VR46266.2020.00031

[35] ALEXIOU E, YANG N Y, EBRAHIMI T. PointXR: a toolbox for visualization and subjective evaluation of point clouds in virtual reality [C]//Twelfth International Conference on Quality of Multimedia Experience (QoMEX). IEEE, 2020: 1 – 6. DOI: 10.1109/QoMEX48832.2020.9123121

[36] WU X J, ZHANG Y, FAN C L, et al. Subjective quality database and objective study of compressed point clouds with 6DoF head-mounted display [J]. IEEE transactions on circuits and systems for video technology, 2021, 31 (12): 4630 – 4644. DOI: 10.1109/TCSVT.2021.3101484

[37] ZHANG J, HUANG W B, ZHU X Q, et al. A subjective quality evaluation for 3D point cloud models [C]//2014 International Conference on Audio, Language and Image Processing. IEEE, 2015: 827 – 831. DOI: 10.1109/ICALIP.2014.7009910

[38] GUTIÉRREZ J, VIGIER T, LE CALLET P. Quality evaluation of 3D objects in mixed reality for different lighting conditions [J]. Electronic imaging, 2020, 32(11): no. 128. DOI: 10.2352/issn.2470-1173.2020.11.hvei-128

[39] TURK G, LEVOY M. Zippered polygon meshes from range images [C]//21st Annual Conference on Computer Graphics and Interactive Techniques. ACM, 1994: 311 – 318. DOI: 10.1145/192161.192241

[40] MPEG-PCC. MPEG point cloud datasets [DB/OL]. (2017-01-15)[2018-05-26]. http://mpegfs. int-evry. fr/MPEG/PCC/DataSets/pointCloud/CfP/datasets. mpeg point cloud datasets

[41] JPEG. JPEG pleno database [DB/OL]. (2016-11-04)[2018-04-12]. http://plenodb.jpeg.org

[42] HUA L, YU M, HE Z Y, et al. CPC-GSCT: visual quality assessment for coloured point cloud based on geometric segmentation and colour transformation [J]. IET image processing, 2022, 16(4): 1083 – 1095. DOI: 10.1049/ipr2.12211

[43] AK A, ZERMAN E, QUACH M, et al. BASICS: broad quality assessment of static point clouds in compression scenarios [EB/OL]. (2023-02-09)[2023-05-06]. https://arxiv.org/abs/2302.04796

[44] LIU Q, YUAN H, HAMZAOUI R, et al. Reduced reference perceptual quality model with application to rate control for video-based point cloud compression [J]. IEEE transactions on image processing, 2021, 30: 6623 – 6636. DOI: 10.1109/TIP.2021.3096060

[45] LIU Q, SU H L, CHEN T X, et al. No-reference bitstream-layer model for perceptual quality assessment of V-PCC encoded point clouds [J]. IEEE transactions on multimedia, 2023, 25: 4533 – 4546. DOI: 10.1109/TMM.2022.3177926

[46] QUACH M, VALENZISE G, DUFAUX F. Improved deep point cloud geometry compression [C]//IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP). IEEE, 2020: 1 – 6. DOI: 10.1109/MMSP48831.2020.9287077

[47] TIAN D, OCHIMIZU H, FENG C, et al. Geometric distortion metrics for point cloud compression [C]//2017 IEEE International Conference on Image Processing (ICIP). IEEE, 2018: 3460 – 3464. DOI: 10.1109/ICIP.2017.8296925

[48] MEKURIA R, BLOM K, CESAR P. Design, implementation, and evaluation of a point cloud codec for tele-immersive video [J]. IEEE transactions on circuits and systems for video technology, 2017, 27(4): 828 – 842. DOI: 10.1109/TCSVT.2016.2543039

[49] MEKURIA R, LI Z, TULVAN C, et al. Evaluation criteria for PCC (point cloud compression): MPEG: MPEG-I 2016/n16332 [S]. 2016

[50] MEYNET G, DIGNE J, LAVOUÉ G. et al: A quality metric for 3D point clouds [C]//Eleventh International Conference on Quality of Multimedia Experience (QoMEX). IEEE, 2019: 1 – 3. DOI: 10.1109/QoMEX.2019.8743313

[51] JAVAHERI A, BRITES C, PEREIRA F, et al. A generalized Hausdorff distance based quality metric for point cloud geometry [C]//Twelfth International Conference on Quality of Multimedia Experience (QoMEX). IEEE, 2020: 1 – 6. DOI: 10.1109/QoMEX48832.2020.9123087

[52] ALEXIOU E, EBRAHIMI T. Towards a point cloud structural similarity metric [C]//2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). IEEE, 2020: 1 – 6. DOI: 10.1109/ICMEW46912.2020.9106005

[53] MEYNET G, NEHMÉ Y, DIGNE J, et al. PCQM: A full-reference quality metric for colored 3D point clouds [C]//Twelfth International Conference on Quality of Multimedia Experience (QoMEX). IEEE, 2020: 1 – 6. DOI: 10.1109/QoMEX48832.2020.9123147

[54] JAVAHERI A, BRITES C, PEREIRA F, et al. Mahalanobis based point to distribution metric for point cloud geometry quality evaluation [J]. IEEE signal processing letters, 2020, 27: 1350 – 1354. DOI: 10.1109/LSP.2020.3010128

[55] YANG Q, MA Z, XU Y L, et al. Inferring point cloud quality via graph similarity [J]. IEEE transactions on pattern analysis and machine intelligence, 2022, 44(6): 3015 – 3029. DOI: 10.1109/TPAMI.2020.3047083

[56] DINIZ R, FARIAS P G, FARIAS M C Q. Multi-distance point cloud quality assessment [C]//2020 IEEE International Conference on Image Processing (ICIP). IEEE, 2020: 3443 – 3447. DOI: 10.1109/ICIP40778.2020.9190956

[57] DINIZ R, FREITAS P G, FARIAS M C Q. Towards a point cloud quality assessment model using local binary patterns [C]//Twelfth International Conference on Quality of Multimedia Experience (QoMEX). IEEE, 2020: 1 – 6. DOI: 10.1109/QoMEX48832.2020.9123076

[58] DINIZ R, FREITAS P G, FARIAS M. A novel point cloud quality assessment metric based on perceptual color distance patterns [C]//IS&T International Symposium on Electronic Imaging Science and Technology 2021, Image Quality and System Performance XVIII. IS&T, 2021. DOI: 10.2352/issn.2470-1173.2021.9.iqsp-256

[59] DINIZ R, FREITAS P G, FARIAS M C Q. Local luminance patterns for point cloud quality assessment [C]//IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP). IEEE, 2020: 1 – 6. DOI: 10.1109/MMSP48831.2020.9287154

[60] DINIZ R, FREITAS P G, FARIAS M C Q. Color and geometry texture descriptors for point-cloud quality assessment [J]. IEEE signal processing letters, 2021, 28: 1150 – 1154. DOI: 10.1109/LSP.2021.3088059

[61] HUA L, JIANG G Y, YU M, et al. BQE-CVP: blind quality evaluator for colored point cloud based on visual perception [C]//2021 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB). IEEE, 2021: 1 – 6. DOI: 10.1109/BMSB53066.2021.9547070

[62] XU Y L, YANG Q, YANG L, et al. EPES: point cloud quality modeling using elastic potential energy similarity [J]. IEEE transactions on broadcasting, 2022, 68(1): 33 – 42. DOI: 10.1109/TBC.2021.3114510

[63] ZHANG Y J, YANG Q, XU Y L. MS-GraphSIM: inferring point cloud quality via multiscale graph similarity [C]//29th ACM International Conference on Multimedia. ACM, 2021: 1230 – 1238. DOI: 10.1145/3474085.3475294

[64] ZHANG Z C, SUN W, MIN X K, et al. No-reference quality assessment for 3D colored point cloud and mesh models [J]. IEEE transactions on circuits and systems for video technology, 2022, 32(11): 7618 – 7631. DOI: 10.1109/TCSVT.2022.3186894

[65] HE Z Y, JIANG G Y, JIANG Z D, et al. Towards a colored point cloud quality assessment method using colored texture and curvature projection [C]// 2021 IEEE International Conference on Image Processing (ICIP). IEEE, 2021: 1444 – 1448. DOI: 10.1109/ICIP42928.2021.9506762

[66] TAO W X, JIANG G Y, JIANG Z D, et al. Point cloud projection and multiscale feature fusion network based blind quality assessment for colored point clouds [C]//29th ACM International Conference on Multimedia. ACM, 2021: 5266 – 5272. DOI: 10.1145/3474085.3475645

[67] LIU Q, YUAN H, SU H L, et al. PQA-net: Deep no reference point cloud quality assessment via multi-view projection [J]. IEEE transactions on circuits and systems for video technology, 2021, 31(12): 4645 – 4660. DOI: 10.1109/TCSVT.2021.3100282

[68] CIGNONI P, ROCCHINI C, SCOPIGNO R. Metro: measuring error on simplified surfaces [J]. Computer graphics forum, 1998, 17(2): 167 – 174. DOI: 10.1111/1467-8659.00236

[69] TIAN D, OCHIMIZU H, FENG C, et al. Evaluation metrics for point cloud compression: ISO/IEC JTC m74008 [S]. 2017,

[70] HE Z Y, JIANG G Y, YU M, et al. TGP-PCQA: texture and geometry projection based quality assessment for colored point clouds [J]. Journal of visual communication and image representation, 2022, 83: 103449. DOI: 10.1016/j.jvcir.2022.103449

[71] TU R W, JIANG G Y, YU M, et al. Pseudo-reference point cloud quality measurement based on joint 2-D and 3-D distortion description [J]. IEEE transactions on instrumentation and measurement, 2023, 72: No. 5019314. DOI: 10.1109/TIM.2023.3290291

[72] VIOLA I, CESAR P. A reduced reference metric for visual quality evaluation of point cloud contents [J]. IEEE signal processing letters, 2020, 27: 1660 – 1664. DOI: 10.1109/LSP.2020.3024065

[73] LIU Y P, YANG Q, XU Y L. Reduced reference quality assessment for point cloud compression [C]//2022 IEEE International Conference on Visual Communications and Image Processing (VCIP). IEEE, 2023: 1 – 5. DOI: 10.1109/VCIP56404.2022.10008813

[74] ZHOU W, YUE G H, ZHANG R Z, et al. Reduced-reference quality assessment of point clouds via content-oriented saliency projection [J]. IEEE signal processing letters, 2023, 30: 354 – 358. DOI: 10.1109/LSP.2023.3264105

[75] SU H L, LIU Q, LIU Y X, et al. Bitstream-based perceptual quality assessment of compressed 3D point clouds [J]. IEEE transactions on image processing, 2023, 32: 1815 – 1828. DOI: 10.1109/TIP.2023.3253252

[76] ZHOU W, YANG Q, JIANG Q P, et al. Blind quality assessment of 3D dense point clouds with structure guided resampling [EB/OL]. (2022-08-31)[2022-09-05]. https://arxiv.org/abs/2208.14603

[77] LIU Q, LIU Y Y, SU H L, et al. Progressive knowledge transfer based on human visual perception mechanism for perceptual quality assessment of point clouds [EB/OL]. (2022-11-30) [2022-12-04]. https://arxiv. org/abs/2211.16646

[78] TU R W, JIANG G Y, YU M, et al. V-PCC projection based blind point cloud quality assessment for compression distortion [J]. IEEE transactions on emerging topics in computational intelligence, 2023, 7(2): 462 – 473. DOI: 10.1109/TETCI.2022.3201619

[79] SHAN Z Y, YANG Q, YE R, et al. GPA-net: no-reference point cloud quality assessment with multi-task graph convolutional network [J]. IEEE transactions on visualization and computer graphics, 2802, 99: 1 – 13. DOI: 10.1109/TVCG.2023.3282802

[80] ZHANG Z C, SUN W, WU H N, et al. GMS-3DQA: projection-based grid mini-patch sampling for 3D model quality assessment [EB/OL]. (2023-06-09) [2023-07-03]. https://arxiv.org/abs/2306.05658

[81] LIU Y, YANG Q, ZHANG Y, et al. Once-training-all-fine: no-reference point cloud quality assessment via domain-relevance degradation description [EB/OL]. (2023-07-04) [2023-07-12]. https://arxiv. org/abs/2307.01567

[82] YANG Q, LIU Y P, CHEN S H, et al. No-reference point cloud quality assessment via domain adaptation [C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2022: 21147 – 21156. DOI: 10.1109/CVPR52688.2022.02050

[83] FAN Y, ZHANG Z C, SUN W, et al. A No-reference quality assessment metric for point cloud based on captured video sequences [C]//24th International Workshop on Multimedia Signal Processing (MMSP). IEEE, 2022: 1 – 5. DOI: 10.1109/MMSP55362.2022.9949359

[84] ZHANG Z C, SUN W, ZHU Y C, et al. Treating point cloud as moving camera videos: a no-reference quality assessment metric [EB/OL]. (2022-09-11) [2022-10-12]. https://arxiv.org/abs/2208.14085

[85] ZHANG Z C, SUN W, MIN X K, et al. MM-PCQA: multi-modal learning for no-reference point cloud quality assessment [EB/OL]. (2022-09-01)[2022-09-27]. https://arxiv.org/abs/2209.00244

[86] LAVOUÉ G, GELASCAE D, DUPONTF, et al. Perceptually driven 3D distance metrics with application to watermarking [C]//Proc. SPIE 6312, Applications of Digital Image Processing XXIX, SPIE. 2006, 6312: 150 – 161. DOI: 10.1117/12.686964

[87] LAVOUÉ G. A multiscale metric for 3D mesh visual quality assessment [J]. Computer graphics forum, 2011, 30(5): 1427 – 1437. DOI: 10.1111/j.1467-8659.2011.02017.x

[88] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: from error visibility to structural similarity [J]. IEEE transactions on image processing, 2004, 13(4): 600 – 612. DOI: 10.1109/TIP.2003.819861

[89] WANG Z, BOVIK A C. Mean squared error: Love it or leave it? A new look at Signal Fidelity Measures [J]. IEEE signal processing magazine, 2009, 26 (1): 98 – 117. DOI: 10.1109/MSP.2008.930649

[90] SUN W, MIN X K, TU D Y, et al. Blind quality assessment for in-the-wild images via hierarchical feature fusion and iterative mixed database training [J]. IEEE journal of selected topics in signal processing, 2023, PP(99): 1 – 15. DOI: 10.1109/JSTSP.2023.3270621

[91] MITTAL A, MOORTHY A K, BOVIK A C. No-reference image quality assessment in the spatial domain [J]. IEEE transactions on image processing, 2012, 21(12): 4695 – 4708. DOI: 10.1109/TIP.2012.2214050

[92] NARVEKAR N D, KARAM L J. A no-reference perceptual image sharpness metric based on a cumulative probability of blur detection [C]//2009 International Workshop on Quality of Multimedia Experience. IEEE, 2009: 87 – 91. DOI: 10.1109/QOMEX.2009.5246972

[93] ZHANG L, ZHANG L, BOVIK A C. A feature-enriched completely blind image quality evaluator [J]. IEEE transactions on image processing, 2015, 24(8): 2579 – 2591. DOI: 10.1109/TIP.2015.2426416

[94] GU K, ZHAI G T, YANG X K, et al. Using free energy principle for blind image quality assessment [J]. IEEE transactions on multimedia, 2015, 17(1): 50 – 63. DOI: 10.1109/TMM.2014.2373812

[95] GU K, ZHAI G T, YANG X K, et al. No-reference image quality assessment metric by combining free energy theory and structural degradation model [C]// 2013 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2013: 1 – 6. DOI: 10.1109/ICME.2013.6607462

[96] MITTAL A, SOUNDARARAJAN R, BOVIK A C. Making a "completely blind" image quality analyzer [J]. IEEE signal processing letters, 2013, 20(3): 209 – 212. DOI: 10.1109/LSP.2012.2227726

[97] ZHANG W X, MA K D, YAN J, et al. Blind image quality assessment using a deep bilinear convolutional neural network [J]. IEEE transactions on circuits and systems for video technology, 2020, 30(1): 36 – 47. DOI: 10.1109/TCSVT.2018.2886771

[98] HARA K, KATAOKA H, SATOH Y. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2018: 6546 – 6555. DOI: 10.1109/CVPR.2018.0068

[99] VQEG. Final report from the video quality experts group on the validation of objective models of video quality assessment: PHASE II [R]. 2003

[100] WANG Z, SIMONCELLI E P, BOVIK A C. Multiscale structural similarity for image quality assessment [C]//Thrity-Seventh Asilomar Conference on Signals, Systems & Computers. IEEE, 2004: 1398 – 1402. DOI: 10.1109/ACSSC.2003.1292216

[101] SHEIKH H R, BOVIK A C. Image information and visual quality [J]. IEEE transactions on image processing, 2006, 15(2): 430 – 444. DOI: 10.1109/TIP.2005.859378

[102] MITTAL A, SAAD M A, BOVIK A C. A completely blind video integrity oracle [J]. IEEE transactions on image processing. IEEE, 2016, 25(1): 289 – 300. DOI: 10.1109/TIP.2015.2502725

[103] SAAD M A, BOVIK A C, CHARRIER C. Blind prediction of natural video quality [J]. IEEE transactions on image processing, 2014, 23(3): 1352 – 1365. DOI: 10.1109/TIP.2014.2299154

[104] KORHONEN J. Two-level approach for no-reference consumer video quality assessment [J]. IEEE transactions on image processing, 2019, 28(12): 5923 – 5938. DOI: 10.1109/TIP.2019.2923051

[105] TU Z Z, WANG Y L, BIRKBECK N, et al. UGC-VQA: benchmarking blind video quality assessment for user generated content [J]. IEEE transactions on image processing. IEEE, 2021, 30: 4449 – 4464. DOI: 10.1109/TIP.2021.3072221

[106] LI D Q, JIANG T T, JIANG M. Quality assessment of in-the-wild videos [C]// 27th ACM International Conference on Multimedia. New York: ACM, 2019: 2351 – 2359. DOI: 10.1145/3343031.3351028

[107] TU Z Z, YU X X, WANG Y L, et al. RAPIQUE: rapid and accurate video quality prediction of user generated content [J]. IEEE open journal of signal processing, 2021, 2: 425 – 440. DOI: 10.1109/OJSP.2021.3090333

[108] SUN W, WANG T, MIN X K, et al. Deep learning based full-reference and no-reference quality assessment models for compressed UGC videos [C]// 2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). IEEE, 2021: 1 – 6. DOI: 10.1109/ICMEW53276.2021.9455999

Biographies

ZHOU Yingjie (zyj2000@sjtu.edu.cn) received his BE degree in electronics and information engineering from China University of Mining and Technology in 2023. He is currently pursuing a PhD degree at the Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, China. His current research interests include 3D quality assessment and virtual digital human.

ZHANG Zicheng (zzc1998@sjtu.edu.cn) received his BE degree from Shanghai Jiao Tong University, China in 2020 and he is currently pursuing a PhD degree at the Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University. His research interest include image quality assessment, video quality assessment, and 3D visual quality assessment.

SUN Wei received his BE degree from the East China University of Science and Technology, China in 2016. He is currently pursuing a PhD degree at the Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, China. His research interests include image quality assessment, perceptual signal processing, and mobile video processing.

MIN Xiongkuo received his BE degree from Wuhan University, China in 2013, and PhD degree from Shanghai Jiao Tong University, China in 2018. From January 2016 to January 2017, he was a visiting student with the University of Waterloo, Canada. From June 2018 to September 2021, he was a postdoctoral researcher with Shanghai Jiao Tong University. From January 2019 to January 2021, he was a visiting postdoctoral researcher with The University of Texas at Austin, USA and the University of Macau, China. He is currently a tenure-track associate professor with the Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University. His research interests include image/video/audio quality assessment, quality of experience, visual attention modeling, extended reality, and multimodal signal processing.

ZHAI Guangtao received his BE and ME degrees from Shandong University, China in 2001 and 2004, respectively, and PhD degree from Shanghai Jiao Tong University, China in 2009. From 2008 to 2009, he was a visiting student with the Department of Electrical and Computer Engineering, McMaster University, Canada, where he was a postdoctoral fellow from 2010 to 2012. From 2012 to 2013, he was a Humboldt Research Fellow with the Institute of Multimedia Communication and Signal Processing, Friedrich Alexander University of Erlangen - Nuremberg, Germany. He is currently a professor with the Department of Electronics Engineering, Shanghai Jiao Tong University. He has published more than 100 journal articles on the topics including visual information acquisition, image processing, and perceptual signal processing.

Spatio-Temporal Context-Guided Algorithm for Lossless Point Cloud Geometry Compression

ZHANG Huiran[1, 2], DONG Zhen[3], WANG Mingsheng[1, 2]

(1. Guangzhou Urban Planning and Design Survey Research Institute, Guangzhou 510060, China;
2. Guangdong Enterprise Key Laboratory for Urban Sensing, Monitoring and Early Warning, Guangzhou 510060, China;
3. State Key Laboratory of Information Engineering in Surveying Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China)

DOI: 10.12142/ZTECOM.202304003

https://kns.cnki.net/kcms/detail/34.1294.TN.20231108.1004.002.html, published online November 8, 2023

Manuscript received: 2023-09-11

Abstract: Point cloud compression is critical to deploy 3D representation of the physical world such as 3D immersive telepresence, autonomous driving, and cultural heritage preservation. However, point cloud data are distributed irregularly and discontinuously in spatial and temporal domains, where redundant unoccupied voxels and weak correlations in 3D space make achieving efficient compression a challenging problem. In this paper, we propose a spatio-temporal context-guided algorithm for lossless point cloud geometry compression. The proposed scheme starts with dividing the point cloud into sliced layers of unit thickness along the longest axis. Then, it introduces a prediction method where both intra-frame and inter-frame point clouds are available, by determining correspondences between adjacent layers and estimating the shortest path using the travelling salesman algorithm. Finally, the few prediction residual is efficiently compressed with optimal context-guided and adaptive fast-mode arithmetic coding techniques. Experiments prove that the proposed method can effectively achieve low bit rate lossless compression of point cloud geometric information, and is suitable for 3D point cloud compression applicable to various types of scenes.

Keywords: point cloud geometry compression; single-frame point clouds; multi-frame point clouds; predictive coding; arithmetic coding

Citation (Format 1): ZHANG H R, DONG Z, WANG M S. Spatio-temporal context-guided algorithm for lossless point cloud geometry compression [J]. ZTE Communications, 2023, 21(4): 17 – 28. DOI: 10.12142/ZTECOM.202304003

Citation (Format 2): H. R. Zhang, Z. Dong, and M. S. Wang, "Spatio-temporal context-guided algorithm for lossless point cloud geometry compression," *ZTE Communications*, vol. 21, no. 4, pp. 17 – 28, Dec. 2023. doi: 10.12142/ZTECOM.202304003.

1 Introduction

With the improvement of multi-platform and multi-resolution acquisition equipment performance, light detection and ranging (LiDAR) technology can efficiently simulate 3D objects or scenes with massive point sets. Compared with traditional multimedia data, point cloud data contain more physical measurement information which represents objects from free viewpoints, even scenes with complex topological structures. This results in strong interactive and immersive effects that provide users with a vivid and realistic visualization experience. Additionally, point cloud data have stronger anti-noise ability and parallel processing capability, which seems to have gained attraction from the industry and academia, notably for application domains such as cultural heritage preservation, 3D immersive telepresence and automatic driving[1 – 2].

However, point cloud data usually contain millions to billions of points in spatial domains, bringing burdens and challenges to the storage space capacity and network transmission bandwidth. For instance, a common dynamic point cloud utilized for entertainment usually comprises roughly one million points per frame, which, at 30 frames per second, amounts to a total bandwidth of 3.6 Gbit/s if left uncompressed[3]. Therefore, the research on high efficiency geometry compression algorithms for point clouds has important theoretical and practical value.

Prior work tackled this problem by directly building grids or on-demand down-sampling, due to limitations in computer computing power and point cloud collection efficiency, which resulted in low spatio-temporal compression performance and loss of geometric attribute feature information. Recent studies were mainly based on computer graphics and digital signal processing techniques to implement block operations on point cloud data[4 – 5] or combined video coding technology[6 – 7] for optimization. In 2017, the Moving Picture Experts Group (MPEG) solicited proposals for point cloud compression and conducted subsequent discussions on how to compress this type of data. With increasing approaches to point cloud compression available and presented, two-point cloud data compression frameworks—TMC13 and TMC2 were issued in 2018. The research above shows remarkable progress has been made in the compression technology of point cloud. However, prior work mostly dealt with the spatial and temporal correlation of point clouds separately but had not yet been exploited to their full potential in point cloud compression.

To address the aforementioned challenges, we introduce a spatio-temporal context-guided method for lossless point cloud geometry compression. We first divide point clouds into unit layers along the main axis. We then design a prediction mode via a travelling salesman algorithm, by adopting spatio-temporal correlation. Finally, the residuals are written into bitstreams with a utilized context-adaptive arithmetic encoder. Our main contributions are as follows.

1) We design a prediction mode applicable to both intra-frame and inter-frame point cloud, via the extended travelling salesman problem (TSP). By leveraging both the spatial and temporal redundancies of point clouds, the geometry prediction can make better use of spatial correlation and therefore enable various types of scenarios.

2) We present an adaptive arithmetic encoder with fast context update, which selects the optimal 3D context from the context dictionary, and suppresses the increase of entropy estimation. As a result, it enhances the probability calculation efficiency of entropy encoders and yields significant compression results.

The rest of this paper is structured as follows. Section 2 gives an outline of related work on point cloud geometry compression. Section 3 firstly presents an overview of the proposed framework. Then, the proposed method is descibed in detail. Experimental results and conclusions are presented in Sections 4 and 5, respectively.

2 Related Work

There have been many point cloud geometry compression algorithms proposed in the literature. CAO et al. [8] and GRAZIOSI et al.[9] conduct an investigation and summary of current point cloud compression methods, focusing on spatial dimension compression technology and MPEG standardization frameworks respectively. We provide a brief review of recent developments in two categories: single-frame point cloud compression and multi-frame point cloud compression.

2.1 Single-Frame Point Cloud Compression

Single-frame point clouds are widely used in engineering surveys, cultural heritage preservation, geographic information systems, and other scenarios. The octree is a widely used data structure to efficiently represent point clouds, which can be compressed by recording information through the occupied nodes. HUANG et al.[10] propose an octree-based method that recursively subdivides the point cloud into nodes with their positions represented by the geometric center of each unit. FAN et al.[11] further improve this method by introducing cluster analysis to generate a level of detail (LOD) hierarchy and encoding it in a breadth-first order. However, these methods can cause distortion due to the approximation of the original model during the iterative process.

To address these limitations, scholars have introduced geometric structure features, such as the triangular surface model[12], the planar surface model[13–14], and the clustering algorithm[15], for inter-layer prediction and residual calculation. RENTE et al.[16] propose a concept of progressive layered compression that first uses the octree structure for coarse-grained encoding and then uses the graph Fourier transform for compression and reconstruction of cloud details. In 2019, MPEG released the geometry-based point cloud compression (G-PCC) technology for both static and dynamic point clouds, which is implemented through coordinate transformation, voxelization, geometric structure analysis, and arithmetic coding step by step[17].

Since certain octants within an octree may be sparsely populated or even empty, some methods have been proposed to optimize the tree structure by pruning sub-nodes and therefore conserve memory allocation. For example, DRICOT et al. [18] propose an inferred direct coding mode (IDCM) for terminating the octree partition based on predefined conditions of sparsity analysis, which involves pruning the octree structure to save bits allocated to child nodes. ZHANG et al. [19] suggest subdividing the point cloud space along principal components and adapting the partition method from the binary tree, quadtree and octree. Compared with the traditional octree partitioning, the hybrid models mentioned above can effectively reduce the number of bits used to represent sparse points, therefore saving nodes that need to be encoded. However, complex hyperparameter conditions and mode determination are required in the process, making it difficult to meet the requirements of self-adaptation and low complexity.

With deep neural networks making significant strides in image and video compression, researchers have explored ways to further reduce bit rates by leveraging super prior guidance and the redundancy of latent space expression during the compression process. QUACH et al.[20] and HUANG et al.[21] propose methods that incorporate these concepts. GUARDA et al. combine convolutional neural networks and autoencoders to exploit redundancy between adjacent points and enhance coding adaptability in Ref. [22]. Recently, WANG et al. [23] propose a point cloud compression method based on the variational auto-encoder, which improves the compression ratio by learning the hyperprior and reducing the memory consumption of arithmetic coding. The aforementioned methods use neural network encoders to capture the high-order hidden vector of the point cloud, the entropy model probabilities, and the edge probabilities of which fit better, thus reducing the memory consumption of arithmetic coding.

Generally speaking, the research on single-frame point cloud geometric compression is relatively mature, but there are two challenges that remain yet. Spatial correlation has not been utilized effectively, and most methods do not code the correlation of point cloud data thoroughly and efficiently. Besides, the calculation of the probability model for entropy coding appears long and arduous due to the massive number of contexts.

2.2 Multi-Frame Point Cloud Compression

Multi-frame point clouds are commonly used in scenarios such as real-time 3D immersive telepresence, interactive VR, 3D free viewpoint broadcasting and automatic driving. Unlike single-frame point cloud compression, multi-frame point cloud compression prioritizes the use of time correlation, as well as motion estimation and compensation. The existing methods for multi-frame point cloud compression can be divided into two categories: 2D projection and 3D decorrelation.

The field of image and video compression is extensive and has been well-explored over the past few decades. Various algorithms convert point clouds into images and then compress them straightforwardly by FFmpeg and H.265 encoders, etc. AINALA et al[24] introduce a planar projection approximate encoding mode that encodes both geometry and color attributes through raster scanning on the plane. However, this method causes changes in the target shape during the mapping process, making accurate inter-prediction difficult. Therefore, SCHWARZ et al.[25] and SEVOM et al.[26] suggest rotated planar projection, cube projection, and patch-based projection methods to convert point clouds into 2D videos, respectively. By placing similar projections in adjacent frames at the same location in adjacent images, the video compressor can fully remove temporal correlation. In Ref. [27], inter-geometry prediction is conducted via TSP, which computes the one-to-one correspondence of adjacent intra-blocks by searching for the block with the closest average value. MPEG released the video-based point cloud compression (V-PCC) technology for dynamic point clouds in 2019[28]. This framework divides the input point cloud into small blocks with similar normal vectors and continuous space, then converts them to the planar surface through cubes to record the occupancy image and auxiliary information. All resulting images are compressed by mature video codecs, and all bitstreams are assembled into a single output file. Other attempts have been made to improve the effectiveness of these methods. COSTA et al.[29] exploit several new patch packaging strategies from the perspective of optimization for the packaging algorithm, data packaging links, related sorting, and positioning indicators. Furthermore, PARK et al. [30] design a data-adaptive packing method that adaptively groups adjacent frames into the same group according to the structural similarity without affecting the performance of the V-PCC stream.

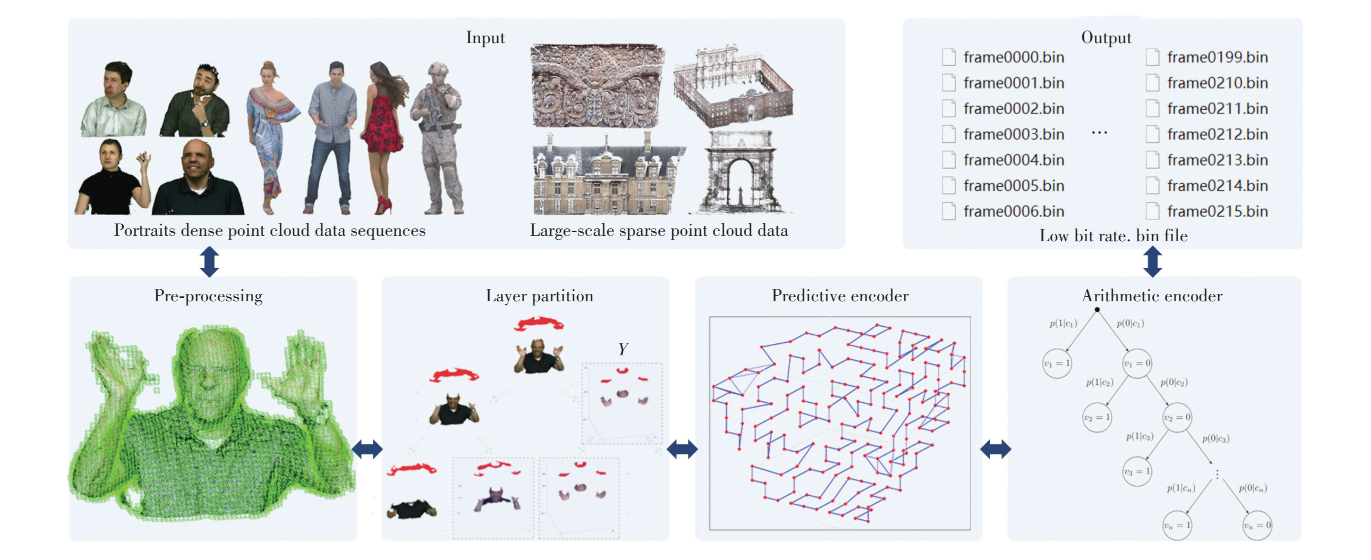
Due to the inevitable information loss caused by point cloud projection, scholars have developed effective techniques to compress the point cloud sequence of consecutive frames using motion compensation technology based on 3D space. KAMMERL et al.[31] propose an octree-based geometric encoding method, which achieves high compression efficiency by performing the exclusive OR (XOR) differences between adjacent frames. This method has been not only adopted in the popular Point Cloud Library (PCL)[32] but also widely used for further algorithm research. Other interframe approaches convert the 3D motion estimation problem into a feature matching problem[33] or use reconstructed geometric information[34] to predict motion vectors and identify the corresponding relationship between adjacent frames accurately. Recent explosive studies[35 – 36] have shown that the learned video compression offers better rate-distortion performance over traditional ones, bringing significant reference significance to point cloud compression. ZHAO et al.[37] introduce a bi-directional inter-frame prediction network to perform inter-frame prediction and bring effective utilization of relevant information in spatial and temporal dimensions. KAYA et al. [38] design a new paradigm for encoding geometric features of dense point cloud sequences, optimizing the CNN for estimating the encoding distribution to realize lossless compression of dense point clouds.

Despite progress in the compression coding technology of multi-frame point cloud models, two problems persist. The existing multi-frame point cloud compression approaches mainly rely on video coding and motion compensation, which inevitably involves information loss or distortion caused by mapping and block edge discontinuity. In addition, predictive coding exhibits low applicability due to the inconsistency of inter-frame point cloud geometry. The apparent offset of points between frames and the unavoidable noise increases the difficulty of effectively using predictive coding in inter-frame compression.

3 Proposed Spatio-Temporal Context-Guided Lossless Geometry Point Cloud Compression Method

3.1 Overview

The overall pipeline of our spatio-temporal context-guided algorithm is shown in Fig. 1. First, we preprocess the input point cloud by applying voxelization and scale transformation. Then, the point cloud is divided into unit thickness sliced layers along the main axis. Next, we design a prediction mode that makes full use of the temporal and spatial correlation information within both intra-frame and inter-frame. We calculate the shortest path of points of reference layers (R-layers) via travelling salesman algorithms, and the results of the R-layers are then used to predict spatio-temporally and encode the rest of the point clouds, namely predicted layers (P-layers). Finally, the improved entropy coding algorithms are adopted to obtain the compressed binary file.



▲Figure 1. Proposed framework for spatio-temporal context-guided lossless point cloud geometry compression

3.2 Image Sliced-Based Hierarchical Division

1) Pre-processing

The pre-processing module includes voxelization and scale transformation, for better indexing of each certain point. In voxelization, we divide the space into cubes of size N, which corresponds to the actual resolution of the point cloud. Each point is assigned a unique voxel based on its position. A voxel is recorded as 1; if it is positively occupied, it is 0 otherwise.

Scale transformation can reduce the sparsity for better compression by zooming out the point cloud, where the distance between points gets smaller. We aggregate the point cloud coordinates (x, y, z) using a scaling factor s, i.e.,

$$\hat{P}_n = P_n \times s = (x_n \times s, y_n \times s, z_n \times s), s \leqslant 1 . \tag{1}$$

To ensure lossless compression, we need to ensure that the scaling factor s cannot cause geometry loss and needs to be recorded in the header file.

2) Sliced-layer division

This module works by dividing the 3D point cloud along one of its axes, creating several unit-sliced layers with occupied and non-occupied information only that can be further compressed using a predictive encoder and an arithmetic coder. The function is defined as:

$$S(a,b) = \text{slice}(G, \text{axis}) = \begin{cases} G(x, a, b),\ \text{if axis} = X \\ G(a, y, b),\ \text{if axis} = Y \\ G(a, b, z),\ \text{if axis} = Z \end{cases}, \tag{2}$$

where G refers to the input point cloud coordinate matrix, axis refers to the selected dimension, and $S(a, b)$ is the 2D slice extracted by each layer.

In general, we conduct experiments on a large number of test sequences, and results suggest that division along the longest axis of point cloud spatial variation yields the lowest bitrate, i.e.

$$\text{axis} = \begin{cases} X, \text{if} \left(x_{\max} - x_{\min}\right) \geqslant \left(y_{\max} - y_{\min}\right), \left(x_{\max} - x_{\min}\right) \geqslant \left(z_{\max} - z_{\min}\right) \\ Y, \text{if} \left(y_{\max} - y_{\min}\right) > \left(x_{\max} - x_{\min}\right), \left(y_{\max} - y_{\min}\right) \geqslant \left(z_{\max} - z_{\min}\right) \\ Z, \text{ if else} \end{cases} . \tag{3}$$

3) Minimum bounding box extraction

In most cases, on-occupied voxels are typically unavoidable and greatly outnumber occupied voxels. As a result, processing and encoding both types of voxels simultaneously burdens the computational complexity and encoding speeds of the compression algorithm. Therefore, we adopt the oriented bounding box (OBB) [39] to calculate the minimum bounding box for each sliced layer, ensuring that the directions of the bounding boxes are consistent across layers. In subsequent processing, only the voxels located within the restricted rectangle are compressed.

3.3 Spatial Context-Guided Predictive Encoding

The goal of spatial context-guided predictive encoding is to encode all the points layer by layer. Inspired by the TSP, we design a prediction mode to explore the potential orders and correlation within each sliced layer. This module consists of partition and the shortest path calculation.

At first, we partition the sliced layers and determine the R-layer and R-layers for each group. We traverse the point cloud layer by layer along the selected axis. When the length of the main direction of the minimum bounding box between adjacent layers differs by a specified unit length, it is recorded as the same group. Otherwise, it is used as the reference layer of the next group, and each point cloud in the following group uses the same shortest path. In this paper, we set the first layer of each group as the R-layer, and the others as P-layers. We also carry out experiments on a large number of test sequences and recommend setting this specified parameter as 3 units to obtain the best compression.

Afterwards, we conduct the shortest path calculation on the R-layers and record the residuals of P-layers. According to the distribution regulation of the point cloud of each slice layer, we optimally arrange the irregular point clouds for each slice layer based on the TSP algorithm. This allows us to efficiently compute the shortest path to the point cloud of the R-layers, and then record the residuals of the corresponding prediction layers. Algorithm 1 shows the pseudo-code of the prediction procedure.

Algorithm 1. Spatial context-guided predictive encoding

```
Input: point cloud sliced-layers
Output: the shortest path $\min \sum_{i,j=1}^{n-1} \mathrm{dist}(\mathrm{pc}_i, \mathrm{pc}_j)$, the shortest path record tables of R-layers, and predictive residuals
Definition: $\mathrm{dist}\left(\mathrm{pc}_i, \mathrm{pc}_j\right) = \mathrm{norm}(\mathrm{pc}_i, \mathrm{pc}_j)$
Initialization: randomly selected point $\mathrm{pc}_1$
while add a new point $\mathrm{pc}_i$ do :
    $\mathrm{path}(P, \mathrm{init}) = min\{\mathrm{path}(P - i, i) + \mathrm{dist}[i][\mathrm{init}]\}, \forall t \in P$
end while
    return $\min \sum_{i,j=1}^{n-1} \mathrm{dist}(\mathrm{pc}_i, \mathrm{pc}_j)$ and shortest path record tables of R-layers
    for P-layers under-process do :
        R-frame $\mathrm{distPC}_i = \min \sum_{i,j=1}^{n-1} \mathrm{dist}(\mathrm{pc}_i, \mathrm{pc}_j)$
        calculate $\mathrm{residuals}_i = \mathrm{diff}(\mathrm{PC}_i(P,:))$
    end for
    return $\mathrm{residuals}_i$
```

Firstly, we define the distance calculation rule between points in the local area and initialize the path state with a randomly selected point pc_1. In each iteration, whenever a new point pc_i is added, the permutation is dynamically updated through the state transition equation $\mathrm{path}(P - i, i)$ until all added points are recorded in P in the order of the shortest path. This process is modified gradually based on the minimal distance criterion. After all iterations are completed in the total shortest path, we calculate the $\min \sum_{i,j=1}^{n-1} \mathrm{dist}(\mathrm{pc}_i, \mathrm{pc}_j)$ in each of the R-layers, and return the shortest path record table of point clouds in each of the R-layers. For further compression, we calculate the deviation of the P-layers from the shortest path of the R-layer within the same group and record them as predictive residuals. Finally, the shortest path of the R-layer and the residuals of each group are output and passed to the entropy encoder to compress prediction residuals further.

3.4 Spatio-Temporal Context-Guided Predictive Encoding

The spatial context-guided prediction mode encodes single-frame point clouds individually. However, applying spatial encoding to each single-frame point cloud separately can miss out on opportunities exposed by the temporal correlations across multi-frame point cloud. Considering that multi-frame point cloud shares large chunks of overlaps, we focus on using temporal redundancy to further enhance the compression efficiency. Hence, based on the proposed spatial context-guided prediction mode, we can compress multi-frame point cloud by identifying a correspondence between adjacent layers across frames.

1) Inter-frame partition

To enhance the effectiveness of inter-frame prediction mode, it is crucial to ensure adequate similarity between adjacent layers of frames. As a result, we need to partition the groups between adjacent frames and determine the R-layers and P-layers across frames. By estimating the shortest path of the P-layers based on the shortest path of the R-layers, we record the prediction residuals and further compress them through the entropy encoder. Algorithm 2 shows the pseudo-code of the inter-frame partition.

Algorithm 2. Inter-frame partition

```
Input: point cloud sliced-layers $S_1,S_2,\cdots,S_n$, and principal axis lengths $h_i$ of $S_i$ inter-frame point cloud sliced layers $SS_1,SS_2,\cdots,SS_n$, and principal axis lengths $hh_i$ of $SS_i$
Output: correspondence and partition of the adjacent layers’ relationship
Initialization: set $S_1$ and $SS_1$ as corresponding layers
for new $S_i$ and $SS_i$ do :
    coarse partition: set $S_i$ and $SS_i$ as corresponding layers
    if $\left|h_i - hh_i\right| \leqslant 3$ :
    fine partition: set $S_i$ and $SS_i$ as corresponding layers
    else if
    compare $\left|h_i - hh_i\right|$,$\left|h_i - hh_{i-1}\right|$, and $\left|h_i - hh_{i+1}\right|$, and pick the minimum
     set the slice layer corresponding to the minimum and $SS_i$ as corresponding layers
     else
     set as a single layer
end for
```

Based on sliced-layers orientation alignment, we realize coarse partition and fine partition successively. For coarse partition, we sort the sliced layers of each frame based on the coordinates corresponding to the division axes, from small to large. As a result, each slice layer of each frame has a unique layer number, allowing us to coarsely partition the slice layers with the same number between adjacent frames. Afterward, we compute the difference between the principal axis lengths of the minimum bounding boxes of adjacent layers with the same number. If this value is less than or equal to a specified length unit, the layers will be partitioned into the same group. Otherwise, we compare the difference in the length of the main direction axis of the minimum bounding box in the corresponding layer of the adjacent frame with the specified layer before and after the number in the adjacent frame. The layer with the smallest difference is then partitioned into the same group. This ensures a fine partition between adjacent layers, and so as to realize the fine partition of the adjacent relationship.

2) Spatio-temporal context-guided prediction mode

Based on the partition, we apply and expand the prediction mode mentioned in Section 3.3. We incorporate inter-frame context in the process, meaning that the first layer of each group, which serves as the R-layer, may not necessarily yield the best prediction result. To fully explore the potential correlation between adjacent layers, we need to expose the optimal prediction mode.

Firstly, we calculate the prediction residuals for each sliced-layer in the current group when used as the R-layer. By comparing the prediction residuals in all cases, we select the R-layer with the smallest absolute residual value as the best prediction mode. For R-layer shortest path calculation, we use the travelling salesman algorithm to compute the shortest path of the R-layers under the best prediction mode. Moreover, we calculate the prediction residuals for each group under their respective best prediction modes. We also record the occupancy length and R-layer information of each group for further compression in subsequent processing.

In the follow-up operation, we use arithmetic coding based on the best context selection for the above information to complete the entire process of the multi-frame point cloud geometry compression algorithm.

3.5 Arithmetic Coding Based on Context Dictionary

The massive amount of context in point cloud significantly burdens the overall compression scheme in terms of arithmetic coding computational complexity. We improve the arithmetic coding from the following two modules. 1) We set up a context dictionary, and select and update the global optimal value according to the entropy estimate, and then 2) we adopt adaptive encoders to efficiently calculate the upper and lower bounds of probabilities.

1) Context dictionary construction

We construct a context dictionary that represents a triple queue, consisting of coordinates of the point cloud at each sliced-layer and the integer representation of its corresponding non-empty context. Thus, we associate the voxels contained in the point cloud with the minimum bounding box of each layer with its non-empty context. To illustrate the construction of the triple queue array of the context dictionary clearly, we give an intuitive explanation in Fig. 2.

For the shaded two squares in Fig. 2, only the context map positions pc$_1$ and pc$_2$ are considered. The context contribution along the x-axis and the y-axis is recorded to the two queues $\mathbb{Q}^{X-}$ and $\mathbb{Q}^{Y-}$ respectively. Thus the context dictionary consists of $\mathbb{Q}^{X-}$ and $\mathbb{Q}^{Y-}$. Queue elements with the same coordinates are integrated into a triplet, the context integer representation of which is computed as the sum of the context contributions of the merged triplet.

Therefore, the context of each voxel can be computed as the sum of the independent contributions of occupied voxels in its context dictionary. This structure helps determine whether a voxel should be added to the context dictionary without tedious matrix lookups, resulting in a significant reduction in computational complexity and runtime.

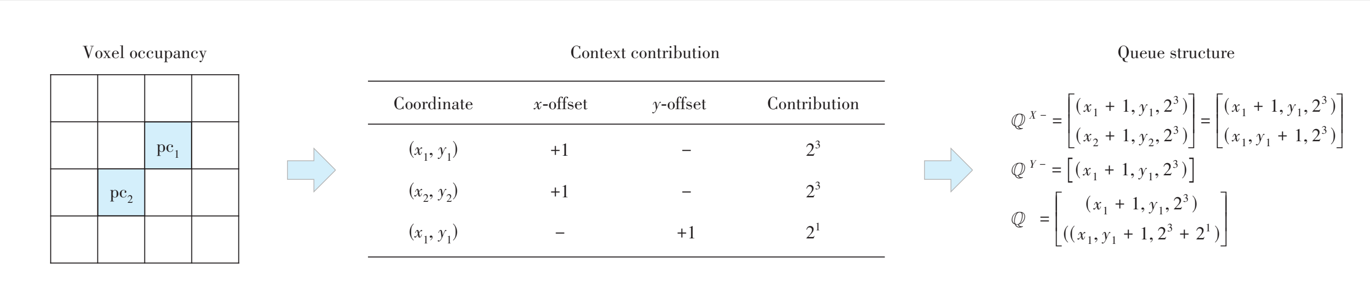
2) Probability calculation

To calculate entropy probability, both the length of the sequence and the context of its constituent voxels must be taken into account. In this module, we design an adaptive encoder that first estimates the upper and lower cumulative probability bounds for each group from the context dictionary, and then encodes it subsequently.

First of all, we construct a binary tree based on the Markov chain model. By traversing the occupancy of voxels, we assign values of 1 and 0 to occupied and empty voxels, respectively, and calculate the probability based on the tree structure. Starting from the root node, when a voxel is occupied, we record the left child node as 1. Otherwise, we mark the right child node as 0 and proceed to the next step of judgment and division. The calculation formula for the run probability of occupied voxels can be found in Eq. (4).

$$P(l) = p\left(1|c_i\right)\cdot\prod_{i-1}^{l-1} p(0|c_i)\,, \tag{4}$$

where l is the length of the run ending at the occupied voxel.



Coordinate	x-offset	y-offset	Contribution
(x_1, y_1)	+1	–	2^3
(x_2, y_2)	+1	–	2^3
(x_1, y_1)	–	+1	2^1

$$\mathbb{Q}^{X-} = \begin{bmatrix} (x_1+1, y_1, 2^3) \\ (x_2+1, y_2, 2^3) \end{bmatrix} = \begin{bmatrix} (x_1+1, y_1, 2^3) \\ (x_1, y_1+1, 2^3) \end{bmatrix}$$

$$\mathbb{Q}^{Y-} = \begin{bmatrix} (x_1+1, y_1, 2^3) \end{bmatrix}$$

$$\mathbb{Q} = \begin{bmatrix} (x_1+1, y_1, 2^3) \\ ((x_1, y_1+1, 2^3+2^1) \end{bmatrix}$$

▲Figure 2. Construction of the context dictionary

For run lengths less than or equal to n, there may be $2n$ of tree nodes representing the occupancy states of voxels. Therefore, the probability of any occupied voxel is represented by the independent joint probability of traversing all states starting at the root and ending at any childless node of the tree.

Based on Eq. (4), to perform arithmetic encoding on the occupancy of the voxel sequence, we need the cumulative upper and lower probabilities of the sequence, as shown in Eq. (5).

$$\begin{cases} \mathrm{Low}(l) = \sum_{r=1}^{l-1} P(r) = \sum_{r=1}^{l-1} p\left(1|c_r\right) \cdot \prod_{i=1}^{r} p(0|c_i) \\ \mathrm{High}(l) = \sum_{r=1}^{l} P(r) = \sum_{r=1}^{l} p\left(1|c_r\right) \cdot \prod_{i=1}^{r} p(0|c_i) \end{cases}. \quad (5)$$

Employing this approach, we can utilize the adaptive properties of arithmetic coding to adjust the probability estimation value of each symbol based on the optimized probability estimation model and the frequency of each symbol in the current symbol sequence. This allows us to calculate the upper and lower bounds of the cumulative probability of occupied voxels and complete the encoding process.

4 Experiment

4.1 Implementation Details

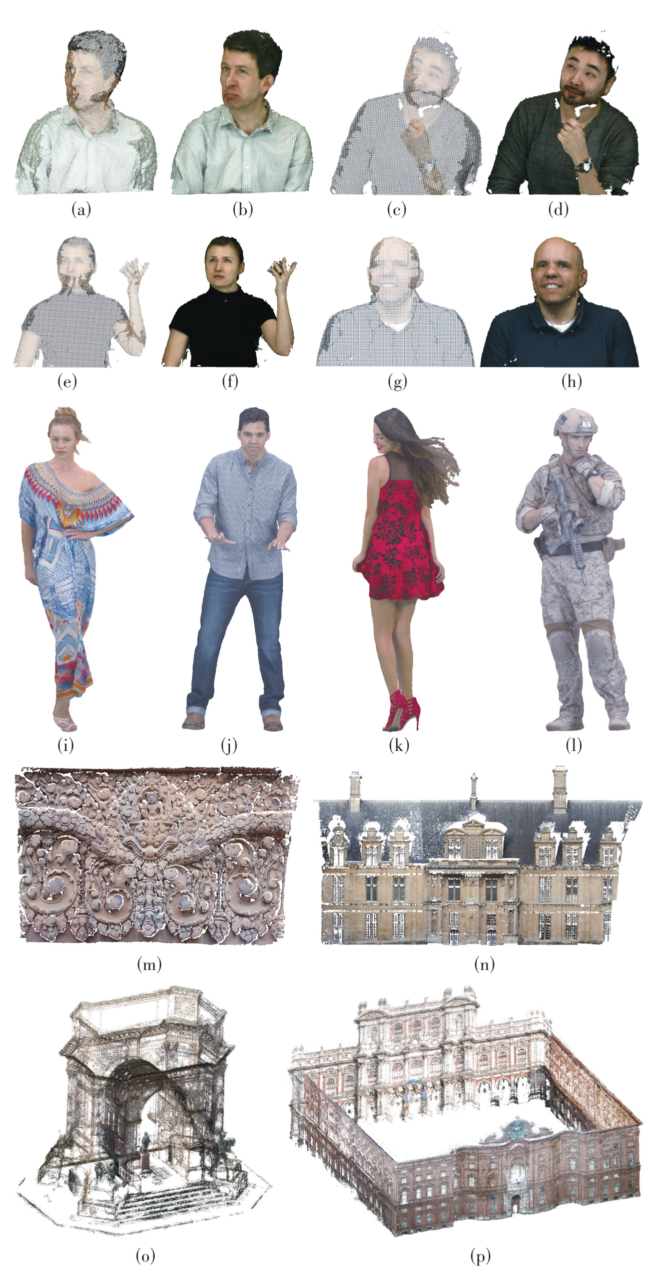
1) Dataset. To verify the performance of our proposed method, extensive experiments were conducted over 16 point cloud datasets that can be downloaded from Ref. [40], as shown in Fig. 3, in which Figs. 3(a) – 3(l) are portraits with dense points, and Figs. 3(m) – 3(p) are architecture with sparse points. Figs. 3(a) – 3(h) are voxelized upper bodies point cloud data sequences of two spatial resolutions obtained from Microsoft. Figs. 3(i) – 3(l) are chosen from 8i voxelized full bodies point cloud data sequences. Remaining large-scale sparse point clouds in Figs. 3(k) – 3(p) are static facade and architecture datasets.

2) Evaluation metrics. The performance of the proposed method is evaluated in terms of bit per point (BPP). The BPP refers to the sum of bits occupied by the coordinate information attached to the point. The lower the value, the better the performance.

$$\mathrm{BPP} = \frac{\mathrm{Size}_{\mathrm{dig}}}{k}, \quad (6)$$

where $\mathrm{Size}_{\mathrm{dig}}$ represents the number of bits occupied by the coordinate information of point cloud data, and k refers to the number of points in the original point cloud.

3) Benchmarks. We mainly compare our method with other baseline algorithms, including: PCL-PCC: octree-based compression in PCL; G-PCC (MPEG intra-coders test model) and interEM (MPEG inter-coders test model) target single-frame and multi-frame point cloud compression respectively; The Silhouette 3D (S3D)[41] and Silhouette 4D (S4D)[42] target single-frame and multi-frame point cloud compression, respectively. For PCL, we use the octree point cloud compression approach in PCL-v1.8.1 for geometry compression only. We set octree resolution parameters from point precision and voxel resolution. For G-PCC (TM13-v11.0), we choose a lossless geometry—lossless attributes condition in an octree-predictive mode, leaving parameters as default. For interEM (tmc3v3.0), we use the experimental results under lossless geometry and



▲ Figure 3. Point cloud sequences used in experiments: (a) Andrew_vox09, (b) Andrew_vox10, (c) David_vox09, (d) David_vox10, (e) Ricardo_vox09, (f) Ricardo_vox10, (g) Sarah_vox09, (h) Sarah_vox10, (i) Longdress_vox10, (j) Loot_vox10, (k) Redandblack_vox10, (l) Soldier_vox10, (m) Facade_00009_vox12, (n) Facade_00015_vox14, (o) Arco_Valentino_Dense_vox12, and (p) Palazzo_Carignano_Dense_vox14

lossless attributes conditions as a comparison[43]. For S3D and S4D, we follow the default conditions and parameters.

4) Hardware. The proposed algorithm is implemented in Matlab and C++ using some functions of the PCL-v1.8.1. All experiments have been tested on a laptop with Intel Core i7-8750 CPU @2.20 GHz with 8 GB memory.

4.2 Results of Single-Frame Point Cloud Compression

1) Compression results of portraits of dense point cloud data sequences

Table 1 shows the performance of our spatial context-guided lossless point cloud geometry compression algorithms compared with PCL-PCC, G-PCC and S3D methods on portraits of dense point cloud data sequences.

It can be seen from Table 1 that for all the point cloud of the same sequences, the proposed method achieves the lowest compression BPP compared with other methods. Our algorithm offers averaged gains from −1.56% to −0.02% against S3D, and gains from −10.62% to −1.45% against G-PCC. It shows a more obvious advantage, that is, the compression performance gains of the proposed algorithm range from −10.62% to −1.45%; For PCL-PCC, the proposed algorithm shows a nearly doubled gain on all sequences, ranging from −154.43% to −85.39%.

2) Compression results of large-scale sparse point cloud data

Because the S3D cannot work in this case, we only compare our spatial context-guided lossless geometry point cloud compression algorithm with PCL-PCC and G-PCC methods on large-scale sparse point cloud data.

Again, our algorithm achieves considerable performance with G-PCC and PCL-PCC, as shown in Table 1. Results have shown that averaged BPP gains ranging from −8.84% to −4.35% are captured compared with G-PCC. For PCL- PCC, our proposed algorithm shows more obvious advantages, with gains ranging from −34.69% to −23.94%.

3) Summary

To provide a more comprehensible comparison of the single-frame point cloud compression results, Table 2 presents the average results between our spatial context-guided compression method and other state-of-the-art benchmark methods. Compared with S3D, our proposed method shows average gains ranging from −0.58% to −3.43%. As for G-PCC and PCL-PCC, the average gains achieve at least −3.43% and −95.03% respectively.

Experimental analysis reveals that our spatial context-guided compression method exceeds current S3D, G-PCC and PCL-PCC by a significant margin. Thus, it can satisfy the lossless compression requirements of point cloud geometry for various scene types, e.g., dense or sparse distributions, and the effectiveness of our method consistently remains.

4.3 Results of Multi-frame Point Cloud Compression

We evaluate our proposed spatial-temporal context-guided point cloud geometry compression algorithm against existing compression algorithms such as S4D, PCL-PCC, G-PCC and interEM. Only portraits of dense point cloud data sequences are used in this experiment. The results are illustrated in

▼Table 1. BPP comparisons of our spatial context-guided compression algorithm and the baseline methods

Point Cloud Data	BPP/bit				Gains		
	Single ↓	G-PCC ↓	PCL-PCC ↓	S3D ↓	G-PCC/%	PCL-PCC/%	S3D/%
Andrew_vox09	1.118 83	1.135 068	2.074 226	1.12	−1.45	−85.39	−0.10
Andrew_vox10	1.010 745	1.104 986	1.952 745	-	−9.32	−93.20	-
David_vox09	1.058 42	1.114 673	2.105 917	1.06	−5.31	−98.97	−0.15
David_vox10	1.028 09	1.090 388	1.974 752	-	−6.06	−92.08	-
Ricardo_vox09	1.037 76	1.081 282	2.046 144	1.04	−4.19	−97.17	−0.22
Ricardo_vox10	0.965 985	1.068 567	1.944 874	-	−10.62	−101.34	-
Sarah_vox09	1.063 19	1.107 865	2.101 666	1.07	−4.20	−97.68	−0.64
Sarah_vox10	1.012 36	1.065 947	1.978 308	-	−5.29	−95.42	-
Longdress_vox10	0.945 35	1.025 244	2.347 862	0.95	−8.45	−148.36	−0.49
Loot_vox10	0.909 825	0.945 36	2.314 874	0.91	−3.91	−154.43	−0.02
Redandblack_vox10	1.014 15	1.082 107	2.400 688	1.03	−6.70	−136.72	−1.56
Soldier_vox10	0.958 515	1.032 572	2.423 025	0.96	−7.73	−152.79	−0.15
Facade 00009 vox12	6.941 5	7.243 8	9.349 4	-	−4.35	−34.69	-
Facade_00015_vox14	7.937 2	8.638 5	10.030 5	-	−8.84	−26.37	-
Arco_Valentino_ Dense_vox12	9.077 9	9.826 4	11.251 4	-	−8.25	−23.94	-
Palazzo_Carignano_ Dense_vox14	7.647 5	8.164 4	9.943 4	-	−6.76	−30.02	-

BPP: bit per point
PCC: point cloud compression
S3D: Silhouette 3D
G-PCC: geometry-based point cloud compression
PCL: Point Cloud Library

▼Table 2. BPP comparison with state-of-the-art algorithms on single-frame point cloud data

Point Cloud Data	Average BPP/bit				Average Gains		
	Single ↓	G-PCC ↓	PCL-PCC ↓	S3D ↓	G-PCC	PCL-PCC	S3D
Microsoft voxelized upper bodies	1.036 923	1.096 097	2.022 329	1.072 5	−5.71%	−95.03%	−3.43%
8i voxelized full bodies	0.956 96	1.021 321	2.371 612	0.962 5	−6.73%	−147.83%	−0.58%
MPEG Facade and architecture	1.158 62	1.198 392	2.336 034	-	−3.43%	−101.62%	-

BPP: bit per point
G-PCC: geometry-based point cloud compression
MPEG: Moving Picture Experts Group
PCC: point cloud compression
PCL: Point Cloud Library
S3D: Silhouette 3D

Table 3. As we can see, after optimizations in prediction mode and arithmetic encoder, the proposed algorithm shows superiority on all test sequences. Specifically, compared with interEM and G-PCC, the proposed algorithm shows significant gains ranging from −51.94% to −17.13% and −46.62% to −5.7%, respectively. Compared with S4D, the proposed algorithm shows robust improvement ranging from −12.18% to −0.33%. As for PCL-PCC, our proposed algorithm has nearly halved over all test sequences.

Furthermore, we summarize the compression results and gains of the proposed method on the portraits dense point cloud data sequences, listed in Table 4. On average, it delivers gains between −11.5% and −2.59% compared with the spatial context-guided point cloud geometry compression algorithm proposed previously. Moreover, it shows superior average gains of − 19% compared with G-PCC, and has achieved an average coding gain of −24.55% compared with interEM. Additionally, compared with S3D and S4D, it gains more than −6.11% and −3.64% on average respectively.

The overall experimental analysis shows that the spatio-temporal context-guided point cloud compression method can make full use of both the spatial and temporal correlation of adjacent layers within intra-frames and inter-frames. We also improve the global context selection and probability model of the arithmetic encoder to obtain a lower bit rate. The proposed method surpasses the performance of state-of-

▼Table 3. Bit per point comparisons of our spatio-temporal context-guided compression algorithm and the baseline methods

Point Cloud Sequences	BPP/bit					Gains			
	Multiple ↓	G-PCC ↓	InterEM ↓	PCL-PCC ↓	S4D ↓	G-PCC/%	InterEM/%	PCL-PCC/%	S4D/%
Andrew_vox09	1.072 25	1.135 068	-	2.074 226	1.08	−5.86	-	−93.45	−0.72
Andrew_vox10	0.972 24	1.104 986	-	1.952 745	-	−13.65	-	−100.85	-
David_vox09	1.046 565	1.114 673	-	2.105 917	1.05	−6.51	-	−101.22	−0.33
David_vox10	1.020 547	1.090 388	-	1.974 752	-	−6.84	-	−93.50	-
Ricardo_vox09	0.982 66	1.081 282	-	2.046 144	1.02	−10.04	-	−108.23	−3.80
Ricardo_vox10	0.954 235	1.068 567	-	1.944 874	-	−11.98	-	−103.81	-
Sarah_vox09	1.028 745	1.107 865	-	2.101 666	1.04	−7.69	-	−104.29	−1.09
Sarah_vox10	1.008 465	1.065 947	-	1.978 308	-	−5.70	-	−96.17	-
Longdress_vox10	0.896 585	1.025 244	1.056 275	2.347 862	0.95	−14.35	−17.81	−161.87	−5.96
Loot_vox10	0.861 815	0.945 36	1.009 412	2.314 874	0.89	−9.69	−17.13	−168.60	−3.27
Redandblack_vox10	0.970 43	1.082 107	1.140 317	2.400 688	1.01	−11.51	−17.51	−147.38	−4.08
Soldier_vox10	0.704 24	1.032 572	1.070 037	2.423 025	0.79	−46.62	−51.94	−244.06	−12.18

G-PCC: geometry-based point cloud compression
PCC: point cloud compression
PCL: Point Cloud Library
S4D: Silhouette 4D

▼Table 4. Bit per point comparison with state-of-the-art algorithms on multi-frame point cloud data

Average BPP/bit							
Point cloud data	Multiple ↓	Single ↓	G-PCC ↓	InterEM ↓	PCL-PCC ↓	S4D ↓	S3D ↓
Microsoft voxelized upper bodies	1.010 713	1.036 923	1.096 097	-	2.022 329	1.047 5	1.072 5
8i voxelized full bodies	0.858 268	0.956 96	1.021 321	1.069 01	2.371 612	0.91	0.962 5
Average Gains							
Point cloud data		Single	G-PCC	interEM	PCL-PCC	S4D	S3D
Microsoft voxelized upper bodies		−2.59%	−8.45%	-	−100.09%	−3.64%	−6.11%
8i voxelized full bodies		−11.50%	−19.00%	−24.55%	−176.33%	−6.03%	−12.14%

G-PCC: geometry-based point cloud compression
PCC: point cloud compression
PCL: Point Cloud Library
S3D: Silhouette 3D
S4D: Silhouette 4D

the-art algorithms, so as to meet the requirements of point cloud geometry lossless compression in multimedia application scenarios such as dynamic portraits.

4.4 Ablation Study

We perform ablation studies on predictive encoding over 8i voxelized full-body point cloud data sequences to demonstrate the effectiveness of the partition. It can be seen from Table 5 that the improvement shows a stable gain of −70% on multi-frame point cloud compression and −60% on single-frame point cloud compression against the non-partition predictive coding.

Next, we perform an ablation experiment on arithmetic coding to demonstrate the effectiveness of the context dictionary. As shown in Table 6, a robust improvement of −33% on multi-frame point cloud compression and that of −41% on single-frame point cloud compression against the arithmetic coding without context dictionary are observed in our method.

▼Table 5. Ablation study on predictive encoding

Point Cloud Data	Partition		Non-Partition		Gains/%	
	Multiple ↓	Single ↓	Multipl ↓	Single ↓	Multiple ↓	Single ↓
Longdress_vox10	0.896 585	0.945 35	1.501 45	1.514 88	−67.46	−60.25
Loot_vox10	0.861 815	0.909 825	1.477 48	1.493 59	−71.44	−64.16
Redandblack_vox10	0.970 43	1.014 15	1.620 92	1.548 96	−67.03	−52.73
Soldier_vox10	0.704 24	0.958 515	1.521 01	1.563 37	−115.98	−63.10

▼Table 6. Ablation study on arithmetic coding

Point Cloud Data	With Context Dictionary		Without Context Dictionary		Gains/%	
	Multiple ↓	Single ↓	Multiple ↓	Single ↓	Multiple ↓	Single ↓
Longdress_vox10	0.896 585	0.945 35	1.279 66	1.489 1	−42.73	−57.52
Loot_vox10	0.861 815	0.909 825	1.272 72	1.364 27	−47.68	−49.95
Redandblack_vox10	0.970 43	1.014 15	1.294 69	1.435 11	−33.41	−41.51
Soldier_vox10	0.704 24	0.958 515	1.112 31	1.374 98	−57.94	−43.45

4.5 Time Consumption

We test the time consumption to evaluate the algorithm complexity and compare the proposed methods with others. The algorithm complexity is analyzed by encoders and decoders independently, listed in Table 7. As we can see, G-PCC, interEM and PCL-PCC can achieve an encoding time of less than 10 s and a decoding time of less than 5 s for portrait dense point cloud data. They also perform well in large-scale sparse point cloud data compared with others. Our proposed algorithms take around 60 s and 15 s to encode and decode portrait sequences, even more on facade and architecture point cloud data. There is a trade-off between bitrates and compression speed. Compared with S3D and S4D, which take hundreds of seconds to encode, our time-consuming method can show superiority.

In summary, the time consumption of our proposed methods is medium among all the compared algorithms but still necessary to be further improved.

5 Conclusions

In this paper, we propose a spatio-temporal context-guided method for lossless point cloud geometry compression. We consider sliced point cloud of unit thickness as the input unit and adopt the geometry predictive coding mode based on the travelling salesman algorithm, which applies to both intra-frame and inter-frame. Moreover, we make full use of the global context information and adaptive arithmetic encoder based on context fast update to achieve lossless compression and decompression results of point clouds. Experimental results demonstrate the effectiveness of our methods and their superiority over previous studies. For future work, we plan to further study the overall complexity of the algorithm, by reducing algorithm complexity to achieve a high-speed compression rate and low bit rate compression results. A low bit rate and real-time/low-delay supported method is highly desired in various types of scenes.

▼Table 7. Time consumption comparison with state-of-the-art algorithms in encoding and decoding

Encoding Time/s							
Point cloud data	Multiple	Single	S4D	S3D	G-PCC	InterEM	PCL-PCC
Microsoft voxelized upper bodies	52.1	64.2	806.03	489.72	3.813	-	2.235
8i voxelized full bodies	56.7	66.9	904.67	640.85	7.105	4.708	3.549
MPEG facade and architecture	-	111.2	-	-	15.37	-	22.4
Decoding Time/s							
Point cloud data	Multiple	Single	S4D	S3D	G-PCC	InterEM	PCL-PCC
Microsoft voxelized upper bodies	13.7	14.4	117.4	74.03	1.031	-	0.809
8i voxelized full bodies	16.3	17.1	194.25	113.95	1.376	4.10	0.922
MPEG facade and architecture	-	22.4	-	-	2.703	-	7.74

G-PCC: geometry-based point cloud compression
MPEG: Moving Picture Experts Group
PCC: point cloud compression
PCL: Point Cloud Library
S3D: Silhouette 3D
S4D: Silhouette 4D

References

[1] MI X X, YANG B S, DONG Z, et al. Automated 3D road boundary extraction and vectorization using MLS point clouds [J]. IEEE transactions on intelligent transportation systems, 2022, 23(6): 5287 – 5297. DOI: 10.1109/TITS.2021.3052882

[2] DONG Z, LIANG F X, YANG B S, et al. Registration of large-scale terrestrial laser scanner point clouds: a review and benchmark [J]. ISPRS journal of photogrammetry and remote sensing, 2020, 163: 327 – 342. DOI: 10.1016/j.isprsjprs.2020.03.013

[3] GRAZIOSI D, NAKAGAMI O, KUMA S, et al. An overview of ongoing point cloud compression standardization activities: video-based (V-PCC) and geometry-based (G-PCC) [J]. APSIPA transactions on signal and information processing, 2020, 9: e13

[4] DE QUEIROZ R L, CHOU P A. Compression of 3D point clouds using a region-adaptive hierarchical transform [J]. IEEE transactions on image processing, 2016, 25(8): 3947 – 3956. DOI: 10.1109/TIP.2016.2575005

[5] BLETTERER A, PAYAN F, ANTONINI M, et al. Point cloud compression using depth maps [J]. Electronic imaging, 2016, 2016(21):1 – 6

[6] MEKURIA R, BLOM K, CESAR P. Design, implementation, and evaluation of a point cloud codec for tele-immersive video [J]. IEEE transactions on circuits and systems for video technology, 2017, 27(4): 828 – 842. DOI: 10.1109/TCSVT.2016.2543039

[7] DE QUEIROZ R L, CHOU P A. Motion-compensated compression of dynamic voxelized point clouds [J]. IEEE transactions on image processing, 2017, 26 (8): 3886 – 3895. DOI: 10.1109/TIP.2017.2707807

[8] CAO C, PREDA M, ZAHARIA T. 3D point cloud compression: a survey [C]//The 24th International Conference on 3D Web Technology. ACM, 2019: 1 – 9. DOI: 10.1145/3329714.3338130

[9] GRAZIOSI D, NAKAGAMI O, KUMA S, et al. An overview of ongoing point cloud compression standardization activities: video-based (V-PCC) and geometry-based (G-PCC) [J]. APSIPA transactions on signal and information processing, 2020, 9(1): e13. DOI: 10.1017/atsip.2020.12

[10] HUANG Y, PENG J L, KUO C J, et al. Octree-based progressive geometry coding of point clouds [C]//The 3rd Eurographics/IEEE VGTC Conference on Point-Based Graphics. IEEE, 2016: 103 – 110

[11] FAN Y X, HUANG Y, PENG J L. Point cloud compression based on hierarchical point clustering [C]//Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. IEEE, 2014: 1 – 7. DOI: 10.1109/APSIPA.2013.6694334

[12] DRICOT A, ASCENSO J. Adaptive multi-level triangle soup for geometry-based point cloud coding [C]//The 21st International Workshop on Multimedia Signal Processing (MMSP). IEEE, 2019: 1 – 6. DOI: 10.1109/MMSP.2019.8901791

[13] HE C, RAN L Q, WANG L, et al. Point set surface compression based on shape pattern analysis [J]. Multimedia tools and applications, 2017, 76(20): 20545 – 20565. DOI: 10.1007/s11042-016-3991-0

[14] IMDAD U, ASIF M, AHMAD M, et al. Three dimensional point cloud compression and decompression using polynomials of degree one [J]. Symmetry, 2019, 11(2): 209. DOI: 10.3390/sym11020209

[15] SUN X B, MA H, SUN Y X, et al. A novel point cloud compression algorithm based on clustering [J]. IEEE robotics and automation letters, 2019, 4(2): 2132 – 2139. DOI: 10.1109/LRA.2019.2900747

[16] DE OLIVEIRA RENTE P, BRITES C, ASCENSO J, et al. Graph-based static 3D point clouds geometry coding [J]. IEEE transactions on multimedia, 2019, 21(2): 284 – 299. DOI: 10.1109/TMM.2018.2859591

[17] ISO. Geometry-based point cloud compression (G-PCC): ISO/IEC 23090-9 [S]. 2021

[18] DRICOT A, ASCENSO J. Hybrid octree-plane point cloud geometry coding [C]//The 27th European Signal Processing Conference (EUSIPCO). IEEE, 2019: 1 – 5

[19] ZHANG X, GAO W, LIU S. Implicit geometry partition for point cloud compression [C]//Proceedings of 2020 Data Compression Conference (DCC). IEEE, 2020: 73 – 82. DOI: 10.1109/DCC47342.2020.00015

[20] QUACH M, VALENZISE G, DUFAUX F. Learning convolutional transforms for lossy point cloud geometry compression [C]//The 2019 IEEE International Conference on Image Processing (ICIP). IEEE, 2019: 4320 – 4324. DOI: 10.1109/ICIP.2019.8803413

[21] HUANG T X, LIU Y. 3D point cloud geometry compression on deep learning [C]//The 27th ACM International Conference on Multimedia. ACM, 2019: 890 – 898. DOI: 10.1145/3343031.3351061

[22] GUARDA A F R, RODRIGUES N M M, PEREIRA F. Point cloud coding: Adopting a deep learning-based approach [C]//Picture Coding Symposium (PCS). IEEE, 2020: 1 – 5. DOI: 10.1109/PCS48520.2019.8954537

[23] WANG J Q, ZHU H, MA Z, et al. Learned point cloud geometry compression [EB/OL]. [2023-09-01]. https://arxiv.org/abs/1909.12037.pdf

[24] AINALA K, MEKURIA R N, KHATHARIYA B, et al. An improved enhancement layer for octree based point cloud compression with plane projection approximation [C]//SPIE Optical Engineering+Applications. SPIE, 2016: 223 – 231. DOI: 10.1117/12.2237753

[25] SCHWARZ S, HANNUKSELA M M, FAKOUR-SEVOM V, et al. 2D video coding of volumetric video data [C]//Picture Coding Symposium (PCS). IEEE, 2018: 61 – 65. DOI: 10.1109/PCS.2018.8456265

[26] FAKOUR SEVOM V, SCHWARZ S, GABBOUJ M. Geometry-guided 3D data interpolation for projection-based dynamic point cloud coding [C]//The 7th European Workshop on Visual Information Processing (EUVIP). IEEE, 2019: 1 – 6. DOI: 10.1109/EUVIP.2018.8611760

[27] KATHARIYA B, LI L, LI Z, et al. Lossless dynamic point cloud geometry compression with inter compensation and traveling salesman prediction [C]//Data Compression Conference. IEEE, 2018: 414. DOI: 10.1109/DCC.2018.00067

[28] ISO. Visual volumetric video-based coding (V3C) and video-based point cloud compression: ISO/IEC 23090-5 [S]. 2021

[29] PARK J, LEE J, PARK S, et al. Projection-based occupancy map coding for 3D point cloud compression [J]. IEIE transactions on smart processing & computing, 2020, 9(4): 293 – 297. DOI: 10.5573/ieiespc.2020.9.4.293

[30] COSTA A, DRICOT A, BRITES C, et al. Improved patch packing for the MPEG V-PCC standard [C]//IEEE 21st International Workshop on Multimedia Signal Processing (MMSP). IEEE, 2019: 1 – 6. DOI: 10.1109/MMSP.2019.8901690

[31] KAMMERL J, BLODOW N, RUSU R B, et al. Real-time compression of point cloud streams [C]//Proceedings of 2012 IEEE International Conference on Robotics and Automation. IEEE, 2012: 778 – 785. DOI: 10.1109/ICRA.2012.6224647

[32] PCL. Point cloud library. [EB/OL]. [2023-09-01]. http://pointclouds.org/

[33] THANOU D, CHOU P A, FROSSARD P. Graph-based compression of dynamic 3D point cloud sequences [J]. IEEE transactions on image processing, 2016, 25(4): 1765 – 1778. DOI: 10.1109/TIP.2016.2529506

[34] LI L, LI Z, ZAKHARCHENKO V, et al. Advanced 3D motion prediction for video based point cloud attributes compression [C]//Data Compression Conference (DCC). IEEE, 2019: 498 – 507. DOI: 10.1109/DCC.2019.00058

[35] ZHAO L L, MA K K, LIN X H, et al. Real-time LiDAR point cloud compression using Bi-directional prediction and range-adaptive floating-point coding [J]. IEEE transactions on broadcasting, 2022, 68(3): 620 – 635. DOI: 10.1109/TBC.2022.3162406

[36] LIN J P, LIU D, LI H Q, et al. M-LVC: Multiple frames prediction for learned video compression [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2020: 3543 – 3551. DOI: 10.1109/CVPR42600.2020.00360

[37] YANG R, MENTZER F, VAN GOOL L, et al. Learning for video compression with hierarchical quality and recurrent enhancement [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2020: 6627 – 6636. DOI: 10.1109/CVPR42600.2020.00666

[38] KAYA E C, TABUS I. Lossless compression of point cloud sequences using sequence optimized CNN models [J]. IEEE access, 2022, 10: 83678 – 83691. DOI: 10.1109/ACCESS.2022.3197295

[39] DING S, MANNAN M A, POO A N. Oriented bounding box and octree based global interference detection in 5-axis machining of free-form surfaces [J]. Computer-aided design, 2004, 36(13): 1281-1294

[40] ALEXIOU E, VIOLA I, BORGES T M, et al. A comprehensive study of the rate-distortion performance in MPEG point cloud compression [J]. APSIPA transactions on signal and information processing, 2019, 8: e27. doi:10.1017/ATSIP.2019.20

[41] PEIXOTO E. Intra-frame compression of point cloud geometry using dyadic

decomposition [J]. IEEE signal processing letters, 2020, 27: 246 – 250. DOI: 10.1109/LSP.2020.2965322

[42] RAMALHO E, PEIXOTO E, MEDEIROS E. Silhouette 4D with context selection: lossless geometry compression of dynamic point clouds [J]. IEEE signal processing letters, 2021, 28: 1660 – 1664. DOI: 10.1109/lsp.2021.3102525

[43] ISO. Common test conditions for G-PCC document N00106: ISO/IEC JTC 1/SC 29/WG 7 MPEG [S]. 2021

Biographies

ZHANG Huiran received her BE and ME degrees in School of Geodesy and Geomatics and State Key Laboratory of Information Engineering in Surveying Mapping and Remote Sensing, both from Wuhan University, China in 2020 and 2023, respectively. She is currently the surveyor of Guangzhou Urban Planning and Design Survey Research Institute, China. Her research interests include point cloud data processing and compression. She participated in several projects related to the field of remote sensing and published one paper in Geomatics and Information Science of Wuhan University.

DONG Zhen (dongzhenwhu@whu.edu.cn) received his BE and PhD degrees in remote sensing and photogrammetry from Wuhan University, China in 2011 and 2018, respectively. He is a professor with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University. His research interests include 3D reconstruction, scene understanding, point cloud processing as well as their applications in intelligent transportation system, digital twin cities, urban sustainable development and robotics. He received over 10 honors from various national and international competitions and published around 60 papers in various journals and conferences.

WANG Mingsheng received his BE degree in College of Computer Science and Technology from Jilin University, China in 2001, and ME degree in School of Computer Science and Engineering from South China University of Technology, China in 2004. He is currently a senior engineer with Guangzhou Urban Planning & Design Survey Research Institute, China. His research interests include computer applications and software, physiography, and surveying. He received over 20 honors from various national competitions and published around 50 papers in various journals and conferences.

Lossy Point Cloud Attribute Compression with Subnode-Based Prediction

YIN Qian[1], ZHANG Xinfeng[2], HUANG Hongyue[1], WANG Shanshe[1], MA Siwei[1]

(1. School of Computer Science, Peking University, Beijing 100871, China；
2. School of Computer Science and Technology, University of the Chinese Academy of Sciences, Beijing 100049, China)

DOI: 10.12142/ZTECOM.202304004

https: //kns.cnki.net/kcms/detail/34.1294.TN.20231122.1025.002.html, published online November 23, 2023

Manuscript received: 2023-10-07

Abstract: Recent years have witnessed that 3D point cloud compression (PCC) has become a research hotspot both in academia and industry. Especially in industry, the Moving Picture Expert Group (MPEG) has actively initiated the development of PCC standards. One of the adopted frameworks called geometry-based PCC (G-PCC) follows the architecture of coding geometry first and then coding attributes, where the region adaptive hierarchical transform (RAHT) method is introduced for the lossy attribute compression. The upsampled transform domain prediction in RAHT does not sufficiently explore the attribute correlations between neighbor nodes and thus fails to further reduce the attribute redundancy between neighbor nodes. In this paper, we propose a subnode-based prediction method, where the spatial position relationship between neighbor nodes is fully considered and prediction precision is further promoted. We utilize some already-encoded neighbor nodes to facilitate the upsampled transform domain prediction in RAHT by means of a weighted average strategy. Experimental results have illustrated that our proposed attribute compression method shows better rate-distortion (R-D) performance than the latest MPEG G-PCC (both on reference software TMC13-v22.0 and GeS-TM-v2.0).

Keywords: point cloud compression; MPEG G-PCC; RAHT; subnode-based prediction

Citation (Format 1): YIN Q, ZHANG X F, HUANG H Y, et al. Lossy point cloud attribute compression with subnode-based prediction [J]. ZTE Communications, 2023, 21(4): 29 – 37. DOI: 10.12142/ZTECOM.202304004

Citation (Format 2): Q. Yin, X. F. Zhang, H. Y. Huang, et al., "Lossy point cloud attribute compression with subnode-based prediction," *ZTE Communications*, vol. 21, no. 4, pp. 29 – 37, Dec. 2023. doi: 10.12142/ZTECOM.202304004.

1 Introduction

Rapid progress in 3D graphic technologies and capture devices has enabled high-precision digital representations of 3D objects or scenes. Point clouds, as one of the mainstream 3D data formats, can effectively indicate points in real-world scenes through 3D geometric coordinates and corresponding attributes (e.g., color, normal and reflectance). Considering its flexible representation properties, point clouds have been widely applied to various fields, such as autonomous driving, free-viewpoint broadcasting, and heritage reconstruction[1]. However, in addition to a huge amount of data, point clouds are non-uniformly sampled in space, which undoubtedly makes it unfeasible to put point clouds into applications with limited bandwidth and storage space[2]. Therefore, it is necessary to investigate efficient point cloud compression (PCC) schemes.

With an increasing demand for point cloud applications, the Moving Picture Expert Group (MPEG) standardization committee started to conduct PCC-dedicated standards and issued a Call for Proposals (CfP) in 2017[3]. After intensive developments involving academic and industrial meetings, two independent point cloud compression frameworks have been adopted to cover a wider range of immersive applications and data types. One called video-based PCC (V-PCC)[4] adopts projection-based strategies combined with video codecs, which aims for handling dense point clouds. Another called geometry-based PCC (G-PCC)[5] is more specifically designed for relatively sparse point clouds by using the octree-based architecture. The octree representation first proposed for PCC in Ref. [6] can build a progressive 3D structure for point clouds. Specifically, by recursively dividing point clouds from the root node to leaf nodes, the connectivity information between points can be exploited among the unorganized point clouds. Moreover, the topological neighbor information makes it easier to implement techniques similar to prediction or transformation in video coding. In the current G-PCC scheme, geometry and attributes are coded sequentially and multiple coding tools can be selected to suit different application scenarios.

This work was supported in part by China Postdoctoral Science Foundation under Grant No. 2022M720234 and in part by the National Natural Science Foundation under Grant Nos. 62071449 and U21B2012.

For the geometry information, in addition to octree coding[7], the triangle soup (Trisoup) coding[8] is used to approximate the surface of point clouds as a complement to the octree decomposition while predictive tree coding[9] is applied to low-delay use cases. In terms of attributes, there are mainly two branches concerning different advantages. The level of details (LODs)-based prediction scheme[10] aims to near-lossless or lossless compression and also deliver spatial scalability to G-PCC. By contrast, the region adaptive hierarchical transform (RAHT) scheme[11] is more suitable for lossy compression with much lower complexity. Note that the attribute coding framework RAHT is our main focus in this paper.

As the mainstream attribute compression scheme, the RAHT was first proposed in Ref. [12] to provide a hierarchical transform structure. In general, the RAHT is an adaptive variant of the Haar wavelet transform (HWT), which can evolve into a 3D version of the Haar transform when all nodes are occupied. To furthermore improve the coding efficiency of the RAHT, an upsampled transform domain prediction[13] was proposed and has been adopted in the current G-PCC. Specifically, decoded attributes of the nodes at lower levels (i. e., lower resolution) are used to predict attributes of the nodes at higher levels. Then, the prediction residuals can be further quantized and entropically encoded. During the transform domain prediction stage, in addition to the nodes at lower levels, the nodes at current encoding levels can also be applied to prediction by means of weighted average[14]. However, the information of surrounding neighbor nodes has not been fully utilized in certain search ranges, which means that further exploring the correlations between neighbor nodes can lead to better attribute compression performance.

In this paper, we propose a subnode-based prediction scheme for point cloud attribute compression, which aims at optimizing the upsampled transform domain prediction in RAHT. Specifically, we first analyze the spatial distribution among neighbor nodes and further explore their reference relationship. Based on this analysis, the prediction accuracy is further improved by exploiting some already-encoded nodes that are not used in the current prediction. Then, a weighted average strategy is introduced for the final attribute prediction of the node to be encoded. Extensive simulations are conducted and compared with the G-PCC as the anchor. Experimental results have confirmed that our proposed method outperforms both Test Model Category 13 (TMC13) and Geometry-Based Solid Content Test Model (GeS-TM) platforms in terms of all point cloud datasets provided by MPEG.

The remainder of this paper is organized as follows. Section 2 succinctly reviews the related works on attribute coding in PCC and describes the current RAHT scheme of G-PCC in particular. Our proposed subnode-based prediction approach is then presented in Section 3. Section 4 provides experimental results and analysis and Section 5 concludes this paper.

2 Related Work

In this section, we first review the attribute coding schemes for PCC. They can be mainly divided into three categories, which are projection-based methods, prediction-based methods and transform-based methods. All of them have been introduced to the MPEG PCC standards. To be more specific, the V-PCC utilizes projection-based methods while the other two strategies have been adopted by the G-PCC. Since our work mainly focuses on the geometry-based PCC, the research related to video-based PCC is outside the scope of this paper. Moreover, the current RAHT scheme and upsampled transform domain prediction of G-PCC are also specifically described as our background.

2.1 Point Cloud Attribute Coding Technologies

Among the existing attribute coding approaches, the prediction-based technology is one of the popular schemes to exploit spatial attribute correlations between points. For example, the attribute prediction framework in G-PCC[10] introduces a linear interpolation process by using the k-nearest neighbors (KNN) search algorithm. This prediction method is based on a LODs structure, which splits the whole point cloud into several subsets (i. e., refinement levels) according to the distance criterion. Based on the LODs, the point clouds are then reordered and encoded, where attributes of points are always predicted by their KNN in the previous LODs. Furthermore, an additional flag is provided in Ref. [15] to allow predictions by using points at the same level. On top of the prediction framework, a lifting scheme[16] is proposed to promote attribute lossy coding. To be more specific, compared with the original prediction method, an update operator combined with an adaptive quantization strategy is added to improve the prediction accuracy. Attributes of points in lower LODs are always assigned much higher influence weights because they are used as reference points with higher frequency and probability for predicting points in higher LODs.

Based on the above two prediction schemes, substantial works are investigated to further improve the attribute compression efficiency. WEI et al. propose an enhanced intra-prediction scheme[17] by considering the overall geometric distribution of the neighbors set. They introduce the centroid-based criterion to measure the distribution uniformity of points in a predictive reference set. Since this scheme predictively encodes the point clouds point by point, the prediction errors will accumulate and propagate, especially for points in higher LODs. Hence, a bilateral filter is proposed in Ref. [18] to update the reconstruction values of decoded points, which reduces error propagation when encoding subsequent points. In addition, YIN et al. attempt to optimize the predictive neighbor set by using the normal of point clouds[19], aiming at improving prediction precision for Light Detection and Ranging (LiDAR) point clouds.

Besides prediction-based methods, other approaches contribute to reducing the attribute spatial redundancy in a transform domain. For instance, to utilize the 2D discrete cosine transform (DCT), ZHANG et al. project the point clouds onto two-dimensional grids for color compression[20]. This 3D-to-2D-based method inevitably fails to fully consider three-dimensional spatial correlations. Hence, 3D-DCT-based methods are developed continuously, such as Refs. [21] and [22].

Apart from DCT, more complex transforms are introduced to attribute coding for PCC. The graph Fourier transform (GFT) is first applied to PCC in Ref. [23], which is an extension of the Karhunen-Loève transform (KLT). The graphs are constructed based on octree-decomposed point clouds, where the graph Laplacian matrix can be deduced by connecting the points within small neighborhoods. Then attributes are transformed, quantized, and entropically encoded. Since the coding efficiency of the graph-based methods outperforms the DCT-based method, extensive follow-up works have been carried out on the attribute graph transform coding. Specifically, an optimized graph transform method[24] is proposed to improve the Laplacian sparsity combined with k-dimensional tree partition and an RDO-based quantization process. Then, XU et al.[25] introduce the normal of point clouds, in addition to geometric distance, to measure the connectivity between neighbor points. Moreover, they propose a predictive generalized graph transform scheme[26] to eliminate the temporary redundancy. Although the graph-based transform approaches exhibit superior coding performance, complicated matrix decomposition leads to real-time difficulties in PCC.

Taking the complexity into consideration, RAHT is proposed in Ref. [12] and finally adopted in G-PCC as the fundamental framework. Our work is closely related to the RAHT and corresponding techniques, which will be concisely described in Section 2.2.

2.2 RAHT in MPEG G-PCC

RAHT is a Haar-inspired method with a hierarchical structure, which can be regarded as an extension of 1D HWT. The core of HWT is to represent functions and signals by using a series of wavelets or basis functions. Specifically, suppose a signal S has N elements. The HWT decomposes the original signal S into low-pass and high-pass components, which can be calculated as follows:

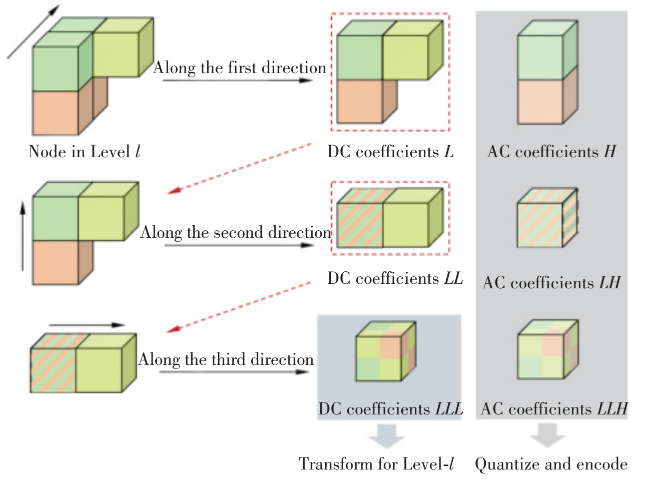
$$\begin{bmatrix} \mathrm{DC} \\ \mathrm{AC} \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} S_{2n} \\ S_{2n+1} \end{bmatrix}, \tag{1}$$

where S_{2n} and S_{2n+1} denote two adjacent elements of signal s. The direct current (DC) and alternating current (AC) coefficients represent the low-frequency and high-frequency parts of the signal respectively. Generally speaking, the energy of the signal after the HWT is mainly concentrated on a few coefficients, especially the DC coefficients, and then appropriate quantization and entropy coding can achieve the purpose of compression.

In order to apply HWT for 3D point cloud attribute compression, 1D HWT is applied sequentially along the x, y, and z directions. Specifically, the RAHT is conducted on a hierarchical octree based on the geometry information of point clouds, which starts from the leaf nodes (i.e., highest resolution level) and proceeds backward until the octree's root node (i.e., lowest resolution level). In each level, the RAHT is applied to each unit node containing 2×2×2 subnodes. As shown in Fig. 1, the unit node is transformed along three directions to generate both DC and AC coefficients, where the DC coefficients along each direction will continue to be transformed while the AC coefficients will be output to be quantized and encoded. Note that the number of coefficients is the same as the number of occupied subnodes in a unit node, including one DC coefficient and several AC coefficients. Then, the DC coefficient obtained from the node at Level l will be used as the attribute of the node at Level l-1 for further transformation. After processing all unit nodes (N occupied nodes) at Level l, N generated DC coefficients (denoted as LLL) continue to be transformed until the root node.

It should be noted that, in the current G-PCC, the dyadic RAHT decomposition[27] is adopted to adapt to more complicated textures. The whole process of the dyadic RAHT is exactly the same as the normal RAHT mentioned above, except that the AC coefficients obtained in each direction will be further transformed like the DC coefficients. Another point to be emphasized is that, unlike HWT in Eq. (1), the wavelet transform kernel for RAHT is modified according to

$$\mathrm{RAHT}\left(w_1, w_2\right) = \frac{1}{\sqrt{w_1 + w_2}} \begin{bmatrix} \sqrt{w_1} & \sqrt{w_2} \\ -\sqrt{w_2} & \sqrt{w_1} \end{bmatrix}, \tag{2}$$



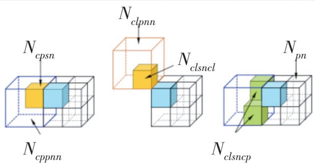
AC: alternating current DC: direct current

▲**Figure 1. Transform procedure of a unit node along three directions. The DC coefficient is denoted as *L* and *H* represents the AC coefficient. *LL* and *LH* represent the DC coefficient and AC coefficient of DC coefficient respectively, and so on** where w_1 and w_2 represent the number of points in two adjacent nodes, which makes the transform more adaptive to the sparsity of point clouds.

To further explore the local spatial correlation, the inter-depth upsampling (IDUS) method[13] is proposed to predict attributes of nodes in the transform domain. As shown in Fig. 2, the upsampling process is realized by means of a weighted average based on geometric distance in the mean attribute space. During the prediction procedure, for each node at Level *l*, there are mainly two types of nodes used for prediction, which are parent-level neighbors at Level *l*-1 and subnode-level neighbors at Level *l*[14]. However, there are still some already-encoded neighbors that are not utilized in the current prediction and the prediction reference relationship can be further refined to improve the attribute compression efficiency.

3 Proposed Approach

Based on the framework described in Section 2.2, we propose a subnode-based transform domain prediction for RAHT that considers more accurate spatial correlations among nodes. In addition to the parent-level neighbors and subnode-level neighbors in G-PCC, some other effective neighbors are also utilized for upsampled transform domain prediction. As illustrated in Fig. 3, the parent-level neighbors at level *l*-1 include three types of nodes, which are the parent node (*Npm*), co-plane parent neighbor node (sharing a side with the subnode to be predicted, *Ncppnn*), and co-line parent neighbor node (sharing an edge with the subnode to be predicted, *Nclpnn)*. For the subnode-level neighbors at Level *l*, there are the co-plane subnode in the co-plane parent neighbor node (*Ncpsn*), co-line subnode in the co-line parent neighbor node (*Nclsncl*) and proposed co-line subnode in the co-plane parent neighbor node (*Nclsncp*). With these predictive reference nodes, we design an optimized transform domain prediction for RAHT. Compared with the original prediction scheme in G-PCC, we first introduce already-encoded neighbor nodes *Nclsncp* as reference candidates and the neighbor search for nodes *Nclsncp* is described in Section 3.1. Since a new type of predictive reference neighbors is added, we further propose a geometric distribution-based prediction to refine the original node prediction reference relationship, which is then detailed in Section 3.2.

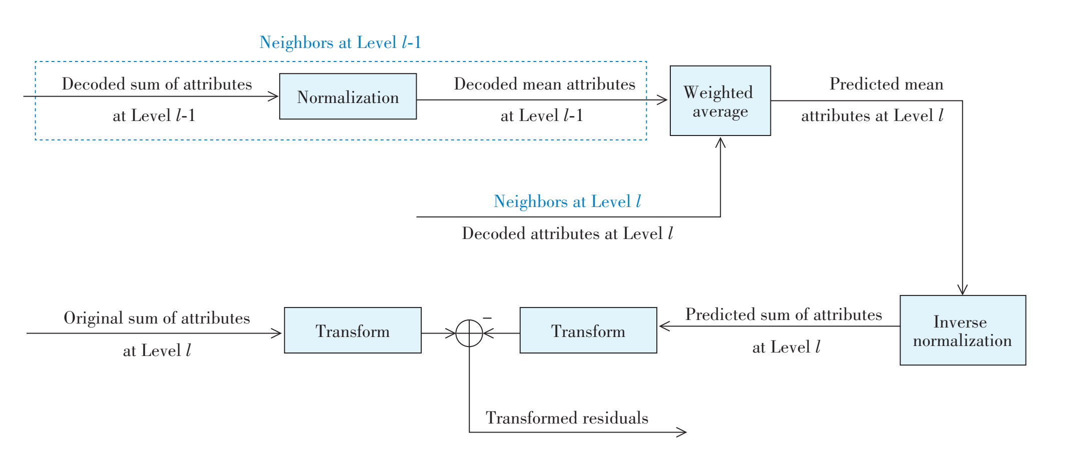
Type	Description	Symbols in Diagram
Parent−level nodes in Level *l*-1	Parent node Co-plane parent neighbor node Co-line parent neighbor node	 N_{clpnn} N_{cpsn} N_{clsncl} N_{pn} N_{cppnn} N_{clsncp}
Subnode−level nodes in Level *l*	Co-plane subnode Co-line subnode in N_{clpnn} **Co-line subnode in N_{cppnn}**	

▲ **Figure 3. Notations of different types of nodes for transform domain prediction, which include parent-level neighbor nodes at level *l*-1 and subnode-level neighbor nodes at level *l***

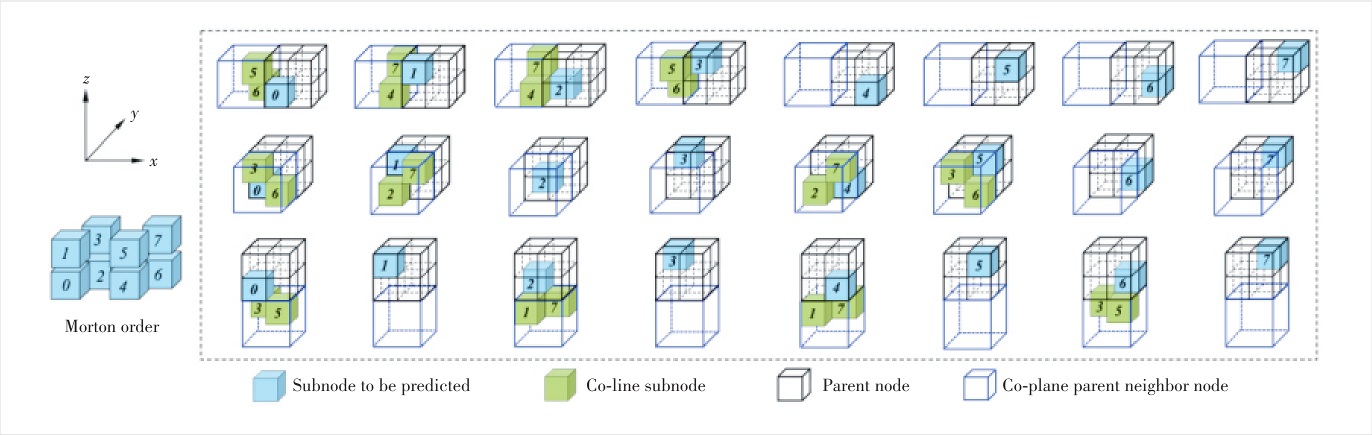
3.1 Neighbor Search for Reference Candidates

Since the proposed co-line subnodes (*Nclsncp*) exist in the co-plane parent neighbor nodes (*Ncppnn*), the neighbor search is mainly decomposed into two stages: 1) determining the co-plane parent neighbor nodes and 2) deciding co-line subnodes reference candidates. Specifically, for each subnode to be predicted, the corresponding parent node has at most six co-plane parent neighbor nodes. Among them, already-encoded subnodes *Nclsncp* can only exist in three parent neighbor nodes (Fig. 4), which are located on the left, front and bottom of *Npn* respectively.

Further considering the position of each subnode to be predicted in its *Npn* as well as the distribution of corresponding *Ncppnn*, the detailed existence of *Nclsncp* of each subnode to be predicted is shown in Fig. 4. Note that the position indexes (from 0 to 8) are organized according to the Morton order. Specifically, we denote *Ntbp i* as the *i*-th subnode to be predicted in the same *Npn*. Then, it can be seen that only *Ntbp* 0 contains six



▲ **Figure 2. Upsampled transform domain prediction for region adaptive hierarchical transform (RAHT) in geometry-based point cloud compression (G-PCC), where upsampling prediction is performed in the mean attribute space and transformation is performed in the sum attribute space**



▲Figure 4. Schematic diagram of co-line subnodes of each subnode to be encoded in the same parent node, where the position indexes are organized according to the Morton order

Nclsncp reference candidates. *Ntbp* 1, *Ntbp* 2 and *Ntbp* 4 include four *Nclsncp* reference candidates. *Ntbp* 3, *Ntbp* 5 and *Ntbp* 6 include two *Nclsncp* reference candidates. *Ntbp* 7 has no *Nclsncp* reference candidate. More detailed information is listed as follows:

- *Ntbp* 0 candidates located in No. 5 and No. 6 subnodes of the left *Ncppnn*, No. 3 and No. 5 subnodes of the bottom *Ncppnn*, and No. 3 and No. 6 subnodes of the front *Ncppnn*.
- *Ntbp* 1 candidates located in No. 4 and No. 7 subnodes of the left *Ncppnn* and No. 2 and No. 7 subnodes of the front *Ncppnn*.
- *Ntbp* 2 candidates located in No. 4 and No. 7 subnodes of the left *Ncppnn* and No. 1 and No. 7 subnodes of the bottom *Ncppnn*.
- *Ntbp* 3 candidates located in No. 5 and No. 6 subnodes of the left *Ncppnn*.
- *Ntbp* 4 candidates located in No. 2 and No. 7 subnodes of the front *Ncppnn* and No. 1 and No. 7 subnodes of the bottom *Ncppnn*.
- *Ntbp* 5 candidates located in No. 3 and No. 6 subnodes of the front *Ncppnn*.
- *Ntbp* 6 candidates located in No. 3 and No. 5 subnodes of the bottom *Ncppnn*.

Among these reference candidates described above, the existing occupied (i. e., non-empty node) *Nclsncp* can be searched by using the relative position relationship with each corresponding *Ntbp i*. In addition to the proposed *Nclsncp*, we also introduce a prediction scheme based on geometric distribution by using *Npn*, *Ncppnn* and *Nclpnn* at Level *l*-1 and *Ncpsn* and *Nclsncl* at Level *l*, which will be detailed in the next section.

3.2 Prediction Based on Geometric Distribution

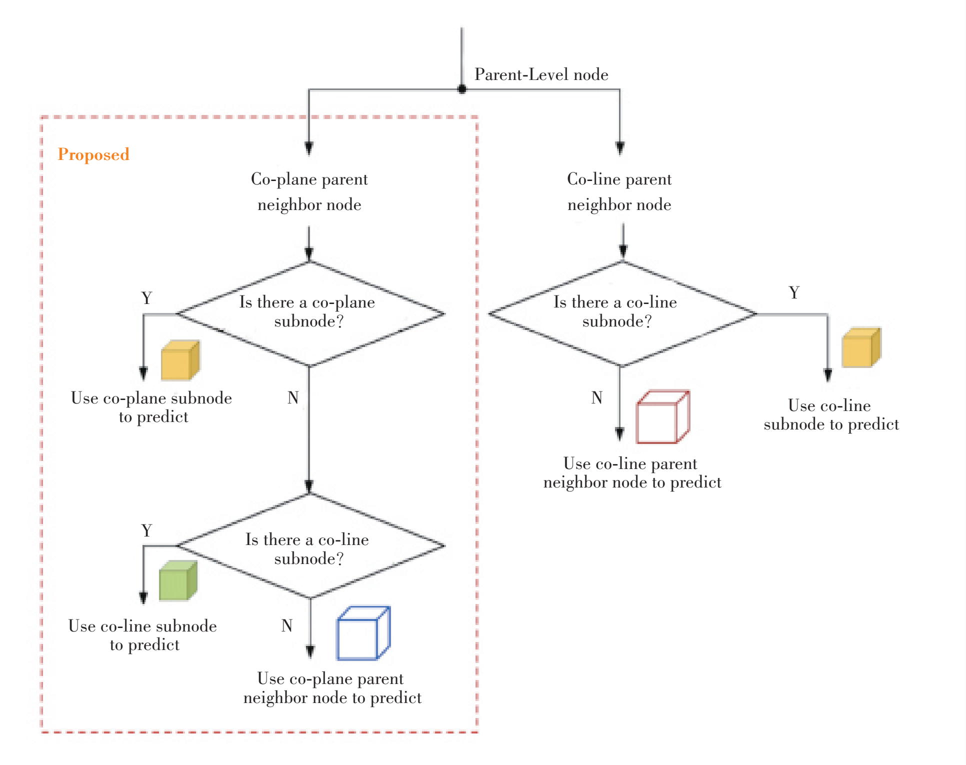
For each subnode to be predicted *Ntbp i* in its parent node *Npn*, we propose to predict them according to the distribution of their neighbor subnodes in the co-plane parent node neighbor nodes. First of all, the distribution can be mainly divided into the following three categories, a total of six subcategories, mainly including:

- Distribution 1: The existing *Ncppnn* contains *Ncpsn*, including three cases: 1) only one *Ncpsn*, 2) one *Ncpsn* and one *Nclsncp*, and 3) one *Ncpsn* and two *Nclsncp*.
- Distribution 2: The existing *Ncppnn* does not contain *Ncpsn* but contains at least one *Nclsncp*, including two cases: 1) only one *Nclsncp* and 2) two *Nclsncp*.
- Distribution 3: The existing *Ncppnn* does not contain any of *Ncpsn* and *Nclsncp*.

Then, the corresponding target prediction mode can be determined by the three types of neighbor subnode distributions. For each subnode to be predicted, in addition to their *Npn* that will definitely participate in the prediction, the prediction reference of other nodes is shown in Fig. 5. Specifically, for *Ncppnn*, we will first determine whether it contains *Ncpsn*, and if so (i.e., satisfying Distribution 1), the attribute value of *Ncpsn* will be used as the prediction instead of the attribute value of its corresponding *Ncppnn* whether it contains *Nclsncp* or not. Then, if there is no *Ncpsn* in *Ncppnn*, we further determine whether it contains at least *Nclsncp*, and if so (i.e., satisfying Distribution 2), the average attribute value of *Nclsncp* will be used as the prediction instead of the attribute value of its corresponding *Ncppnn*. If it contains neither of the above two conditions (i.e., satisfying Distribution 3), the attribute value of *Ncppnn* will be directly used for prediction. Besides *Ncppnn*, for *Ncppnn*, the attribute value of *Nclsncl* will be used as the prediction instead of the attribute value of its corresponding *Nclpnn* if it has *Nclsncl*, which is the same as the current G-PCC.

4 Experiments

To validate the effectiveness of the proposed method, we implement our subnode-based prediction scheme on top of the latest MPEG G-PCC reference software TMC13-v22.0[28] and GeS-TM-v2.0[29]. Extensive simulations have been conducted in accordance with the common test conditions (CTCs)[30]


Parent-Level node
Proposed
Co-plane parent neighbor node
Co-line parent neighbor node
Is there a co-plane subnode?
Is there a co-line subnode?
Y
N
Use co-plane subnode to predict
Use co-line parent neighbor node to predict
Use co-line subnode to predict
Is there a co-line subnode?
Y
N
Use co-line subnode to predict
Use co-plane parent neighbor node to predict

▲**Figure 5. Transform domain prediction based on the three types of neighbor subnode geometric distributions**

where the octree and RAHT configuration are applied for geometry and attribute respectively. In terms of the test conditions, as shown in Table 1, C1 (i.e., lossless geometry lossy attributes) and C2 conditions (i.e., near-lossless/lossy geometry lossy attributes) are both evaluated on the reference software TMC13-v22.0 and GeS-TM-v2.0.

▼**Table 1. Common test conditions in G-PCC**

G-PCC Platform	Conditions		Datasets		
	C1	C2	Cat1	Cat2	Cat3
TMC13	√	√	√		√
GeS-TM	√	√		√	

G-PCC: geometry-based point cloud compression

4.1 Datasets

The test datasets provided by MPEG G-PCC can be mainly classified into three categories: Category 1—Static Objects and Scenes datasets (i. e., Cat1), Category 2—Dynamic Objects datasets (i. e., Cat2), and Category 3—Dynamic Acquisition datasets (i. e., Cat3). Specifically, sequences in Cat1 are further divided based on the density and surface continuity of point clouds (i. e., Solid, Dense, Sparse, and Scant). For Cat2, test classes A, B and C indicate the complexity of point clouds, where A is the lowest and C is the highest. The division of Cat3 is more detailed, including automotive frame-based data acquired by spinning and non-spinning LiDAR sensors (i.e., Am-frame) and automotive LiDAR acquired data after fused and reprocessed (i.e., Am-fused). Note that Am-fused datasets have both color and reflectance attributes. In terms of the CTCs, the Cat1 and Cat3 datasets are tested on TMC13v22.0 while the Cat2 datasets are tested on GeS-TMv2.0. All test sequences mentioned above are available in the MPEG content repository[31].

4.2 Performance Evaluations

The attribute compression performances compared with TMC13-v22.0 are shown in Table 2, where the negative Bjontegaard delta (BD) rate illustrates the coding gains against the anchor. From Table 2, it can be seen that consistent coding

▼**Table 2. Performance of the proposed method against TMC13-v22.0 under C1 and C2 configurations**

Dataset Category	C1 End-to-End BD-Attribute Rate/%				C2 End-to-End BD-Attribute Rate/%			
	Luma	Chroma Cb	Chroma Cr	Reflectance	Luma	Chroma Cb	Chroma Cr	Reflectance
Solid average	−0.4	−0.3	−0.4	/	−0.2	−0.3	−0.2	/
Dense average	−0.2	−0.2	−0.2	/	−0.2	−0.5	−0.1	/
Sparse average	−0.2	−0.2	−0.1	/	−0.2	−0.1	−0.3	/
Scant average	−0.2	−0.2	−0.3	/	−0.2	−0.3	−0.2	/
Am-fused average	−0.3	−1.2	−1.1	−1.1	−0.1	−0.6	−0.7	−0.2
Am-frame spinning average	/	/	/	−0.3	/	/	/	−0.2
Am-frame non-spinning average	/	/	/	−0.6	/	/	/	−0.2
Overall average	**−0.2**	**−0.3**	**−0.3**	**−0.5**	**−0.2**	**−0.3**	**−0.2**	**−0.2**
Average encoding/decoding time/%	102/103				100/107			

BD: Bjontegaard delta

gains can be achieved both under C1 and C2 conditions for all categories. Specifically, 0.2%, 0.3% and 0.5% bitrate reduction for luma, chorma and reflectance are obtained under the C1 condition respectively. 0.2%, 0.3% and 0.2% bitrate reduction for luma, chroma Cb and Cr as well as 0.2% bitrate reduction for reflectance are obtained under C2 condition respectively. Especially for Am-fused datasets, there are over 1% coding gains under the C1 condition for chroma Cb (1.2%), chroma Cr (1.1%) and reflectance (1.1%). Besides the R-D performance, the computational complexity is evaluated by using the average encoding and decoding time. There are only 2% and 3% extra increases on the encoding and decoding time on the C1 condition, with no complexity increase for encoding on the C2 condition.

Apart from TMC13-v22.0, we also compare our proposed method with GeSTM-v2.0 for Cat2. As shown in Table 3, consistent coding gains can be also achieved both under C1 and C2 conditions for all Cat2 datasets. Specifically, 0.4%, 0.4% and 0.5% bitrate reduction for luma, chroma Cb and chroma Cr are obtained under the C1 condition respectively. 0.3%, 0.3% and 0.4% bitrate reduction for luma, chroma Cb and Cr are obtained under the C2 condition respectively. In terms of computational complexity, the encoding time increases by 6% while the decoding time increases by 10%.

To further evaluate the prediction effect of the proposed optimization scheme, we also count the errors during the transform domain prediction stage. Specifically, for each sequence in Cat1, prediction errors of all slices are accumulated if the upsampled prediction is enabled. As illustrated in Fig. 6, the average prediction errors of each type of point cloud are all smaller than that of the original prediction scheme in TMC13-v22.0. Therefore, our proposed method can effectively improve compression efficiency by reducing prediction errors.

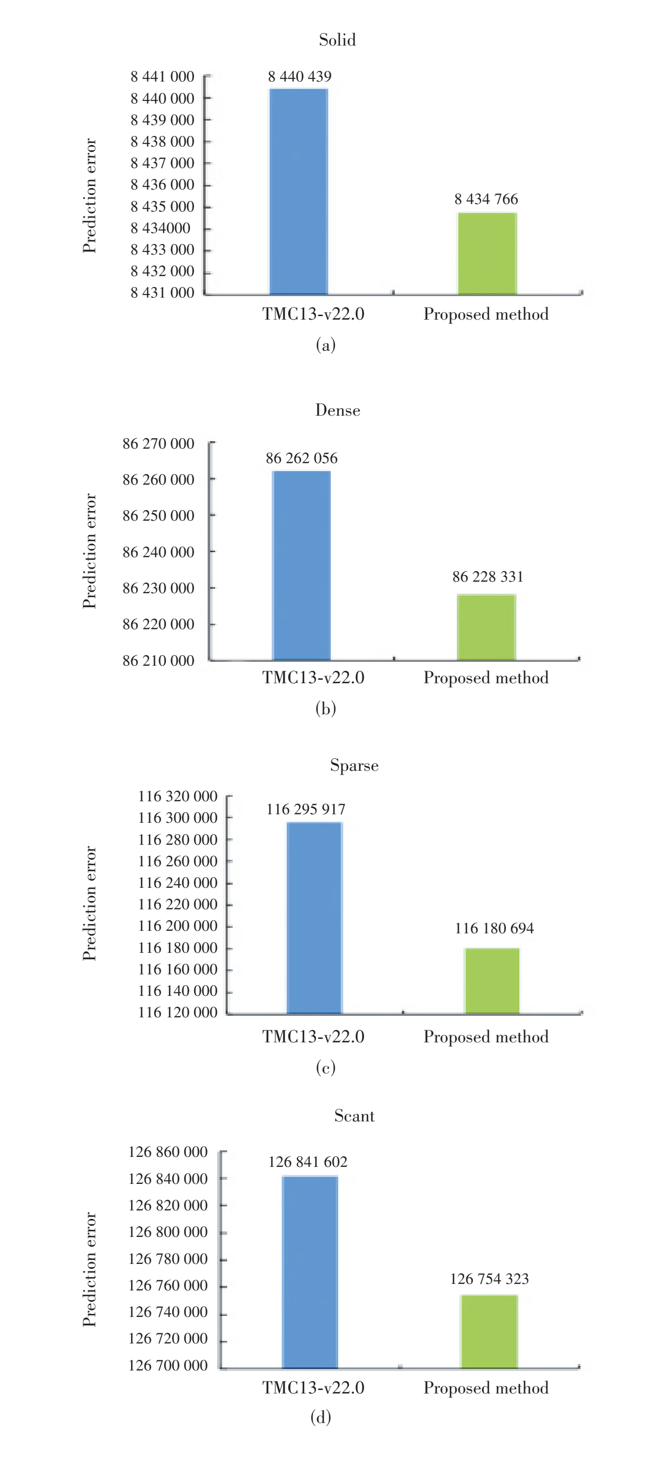
5 Conclusions

In this paper, a subnode-based prediction is proposed to improve the lossy point cloud attribute compression for the MPEG G-PCC platform. Based on the original upsampled transform domain prediction scheme, we leverage some already-encoded neighbor nodes at the same level as the current node to be encoded to optimize the original prediction

▼Table 3. Performance of the proposed method against GeSTM-v2.0 under C1 and C2 configurations

Dataset Category	C1 BD-Rate/%			C2 BD-Rate/%		
	L	Cb	Cr	L	Cb	Cr
Cat2-A average	−0.4	−0.5	−0.5	−0.3	−0.3	−0.4
Cat2-B average	−0.3	−0.3	−0.3	−0.2	−0.2	−0.2
Cat2-C average	−0.5	−0.4	−0.5	−0.4	−0.4	−0.3
Overall average	**−0.4**	**−0.4**	**−0.5**	**−0.3**	**−0.3**	**−0.4**
Avgerage encoding/ decoding time (%)	106/109			101/110		

BD: Bjontegaard delta



▲Figure 6. Prediction errors of the proposed method compared with the original prediction scheme in geometry-based point cloud compression (G-PCC) (i.e., reference software TMC13-v22.0) process. Additionally, a more refined prediction reference relationship is introduced based on the geometric distribution among neighbor nodes. Extensive simulation results demonstrate that our proposed method can achieve consistent coding gains on all types of point clouds, whether sparse LiDAR point clouds, dense colored point clouds, or multi-attribute point clouds, compared to the latest G-PCC test models.

References

[1] TULVAN C, MEKURIA R, LI Z. Use cases for point cloud compression (PCC), output document N16331 [R]. Geneva, Switzerland: ISO/IEC JTC 1/SC29/WG 11 MPEG, 2016

[2] MEKURIA R, BLOM K, CESAR P. Design, implementation, and evaluation of a point cloud codec for tele-immersive video [J]. IEEE transactions on circuits and systems for video technology, 2017, 27(4): 828 – 842. DOI: 10.1109/TCSVT.2016.2543039

[3] MPEG 3D Graphics Coding Group. Call for proposals for point cloud coding v2, output document N16763 [R]. 2017

[4] MPEG 3D Graphics and Haptics Coding Group. V-PCC test model v22, output document N00572 [R]. Antalya, Turkish: ISO/IEC JTC 1/SC 29/WG 11 MPEG, 2023

[5] MPEG 3D Graphics and Haptics Coding Group. G-PCC test model v22, output document N00571 [R]. Antalya, Turkish: ISO/IEC JTC 1/SC 29/WG 11 MPEG, 2023

[6] SCHNABEL R, KLEIN R. Octree-based point-cloud compression [C]//Symposium on Point-Based Graphics. Eurographics Association, 2006: 111 – 120. DOI: 10.2312/SPBG/SPBG06/111-120

[7] PENG J L, KUO C C J. Progressive geometry encoder using octree-based space partitioning [C]//2004 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2004: 1 – 4. DOI: 10.1109/ICME.2004.1394110

[8] NAKAGAMI O. Report on triangle soup decoding, input document m52279 [R]. Brussels, Belgium: ISO/IEC JTC 1/SC 29/WG 11 MPEG, 2020

[9] FLYNN D, TOURAPIS A, MAMMOU K. Predictive geometry coding, input document m51012 [R]. Geneva, Switzerland: ISO/IEC JTC 1/SC 29/WG11 MPEG, 2019

[10] MAMMOU K. PCC test model category 3 v0, output document N17249 [R]. Macau, China: ISO/IEC JTC 1/SC 29/WG 11 MPEG, 2017

[11] CHOU P A, DE QUEIROZ R. L. Transform coder for point cloud attributes, input document m38674 [R]. Geneva, Switzerland: ISO/IEC JTC 1/SC29/WG 11 MPEG, 2016

[12] DE QUEIROZ R L, CHOU P A. Compression of 3D point clouds using a region-adaptive hierarchical transform [J]. IEEE transactions on image processing, 2016, 25(8): 3947 – 3956. DOI: 10.1109/TIP.2016.2575005

[13] FLYNN D, LASSERRE S. G-PCC CE13.18 report on upsampled transform domain prediction in RAHT, input document m49380 [R]. Gothenburg, Sweden: ISO/IEC JTC 1/SC 29/WG 11 MPEG, 2019

[14] WANG W, XU Y, ZHANG Ket al. Sub-node-based prediction in transform domain for RAHT, input document m60203 [R]. Online: ISO/IEC JTC 1/SC 29/WG 11 MPEG, 2022

[15] MAMMOU K. Point cloud compression core experiment 13.6 on attributes prediction strategies, output document N18007 [R]. Macau, China: ISO/IEC JTC 1/SC 29/WG 11 MPEG, 2018

[16] MAMMOU K, TOURAPIS A, KIM J, et al. Proposal for improved lossy compression in TMC1, input document m42640 [R]. San Diego, United states: ISO/IEC JTC 1/SC 29/WG 11 MPEG, 2018

[17] WEI H L, SHAO Y T, WANG J, et al. Enhanced intra prediction scheme in point cloud attribute compression [C]//2019 IEEE Visual Communications and Image Processing (VCIP). IEEE, 2019: 1 – 4. DOI: 10.1109/VCIP47243.2019.8966001

[18] YEA S. VOSOUGHI A, LIU S. Bilateral filtering for predictive transform in G-PCC, input document m46365 [R]. Marrakech, Morocco: ISO/IEC JTC1/SC 29/WG 11 MPEG, 2019

[19] YIN Q, REN Q S, ZHAO L L, et al. Lossless point cloud attribute compression with normal-based intra prediction [C]//2021 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB). IEEE, 2021: 1 – 5. DOI: 10.1109/BMSB53066.2021.9547021

[20] ZHANG X M, WAN W G, AN X D. Clustering and DCT based color point cloud compression [J]. Journal of signal processing systems, 2017, 86(1): 41 – 49. DOI: 10.1007/s11265-015-1095-0

[21] COHEN R A, TIAN D, VETRO A. Point cloud attribute compression using 3-D intra prediction and shape-adaptive transforms [C]//2016 Data Compression Conference (DCC). IEEE, 2016: 141 – 150. DOI: 10.1109/DCC.2016.67

[22] WANG L J, WANG L Y, LUO Y T, et al. Point-Cloud compression using data independent method—a 3D discrete cosine transform approach [C]//2017 IEEE International Conference on Information and Automation (ICIA). IEEE, 2017: 1 – 6. DOI: 10.1109/icinfa.2017.8078873

[23] ZHANG C, FLORÊNCIO D, LOOP C. Point cloud attribute compression with graph transform [C]//2014 IEEE International Conference on Image Processing (ICIP). IEEE, 2014: 2066 – 2070. DOI: 10.1109/ICIP.2014.7025414

[24] SHAO Y T, ZHANG Z B, LI Z, et al. Attribute compression of 3D point clouds using Laplacian sparsity optimized graph transform [C]//2017 IEEE Visual Communications and Image Processing (VCIP). IEEE, 2017: 1 – 4. DOI: 10.1109/VCIP.2017.8305131

[25] XU Y Q, HU W, WANG S S, et al. Cluster-based point cloud coding with normal weighted graph Fourier transform [C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018: 1753 – 1757. DOI: 10.1109/ICASSP.2018.8462684

[26] XU Y Q, HU W, WANG S S, et al. Predictive generalized graph Fourier transform for attribute compression of dynamic point clouds [J]. IEEE transactions on circuits and systems for video technology, 2021, 31(5): 1968 – 1982. DOI: 10.1109/TCSVT.2020.3015901

[27] HOODA R, PAN W D. Early termination of dyadic region-adaptive hierarchical transform for efficient attribute compression of 3D point clouds [J]. IEEE signal processing letters, 2022, 29: 214 – 218. DOI: 10.1109/LSP.2021.3133204

[28] MPEG 3D Graphics and Haptics Coding Group. TMC13 software repository [EB/OL]. (2023-06-02)[2023-11-20]. http://mpegx.int-evry.fr/software/MPEG/PCC/TM/mpeg-pcc-tmc13/-/tags/release-v22.0-rc1

[29] MPEG 3D Graphics and Haptics Coding Group. GeS-TM software repository [EB/OL]. (2023-06-05)[2023-11-20]. http://mpegx.int-evry.fr/software/MPEG/PCC/TM/mpeg-pcc-ges-tm/-/tree/ges-tm-v2.0-rc2

[30] MPEG 3D Graphics and Haptics Coding Group. Common test conditions for G-PCC, output document N00578 [R]. Antalya, Turkish: ISO/IEC JTC 1/SC29/WG 11 MPEG, 2023

[31] MPEG 3D Graphics and Haptics Coding Group. MPEG content repository [EB/OL]. (2023-03-17)[2023-11-20]. http://mpegfs.int-evry.fr/mpegcontent

Biographies

YIN Qian received her MS degree in signal and information processing from University of Electronic Science and Technology of China in 2021. She is currently pursuing a PhD degree in computer science at Peking University, China. She is actively participating in the research work of the Audio Video Coding Standard (AVS) Workgroup of China and Moving Picture Experts Group (MPEG). Her research interests include video and point cloud compression.

ZHANG Xinfeng received his BS degree in computer science from Hebei University of Technology, China in 2007 and PhD degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences in 2014. From 2014 to 2017, he was a research fellow with the Rapid-Rich Object Search

Lab, Nanyang Technological University, Singapore. From October 2017 to October 2018, he was a postdoctoral fellow with the School of Electrical Engineering System, University of Southern California, Los Angeles, USA. From December 2018 to August 2019, he was a research fellow with the Department of Computer Science, City University of Hong Kong, China. He is currently an assistant professor with the School of Computer Science and Technology, University of Chinese Academy of Sciences. He has authored more than 150 refereed journal/conference papers. His research interests include video compression and processing, image/video quality assessment, and 3D point cloud processing.

HUANG Hongyue received his BE degree in communication engineering from Beijing University of Posts and Telecommunications, China in 2015, MS degree in computer science from the Technical University of Berlin (TUB), Germany in 2018, and PhD degree in computer science from the Free University of Brussels (VUB), Belgium in 2021. He is currently a postdoctoral researcher with the National Engineering Research Center of Visual Technology, Peking University, China. His research interests include inventing and optimizing deep-learning-based compression methods for 2D images/videos and 3D visual content such as immersive videos, point clouds, and light field images.

WANG Shanshe received his BS degree from the Department of Mathematics, Heilongjiang University, China in 2004, MS degree in computer software and theory from Northeast Petroleum University, China in 2010, and PhD degree in computer science from the Harbin Institute of Technology, China. He held a postdoctoral position with Peking University, China from 2016 to 2018. He is currently an associate researcher with the School of Electronics Engineering and Computer Science, Institute of Digital Media, Peking University. His current research interests include video compression and image and video quality assessment.

MA Siwei (swma@stu.pku.edu.cn) received his BS degree from Shandong Normal University, China in 1999, and PhD degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences in 2005. He held a postdoctoral position with the University of Southern California, Los Angeles, USA from 2005 to 2007. He is currently a professor with the School of Electronics Engineering and Computer Science, Institute of Digital Media, Peking University, China. He has authored over 300 technical articles in refereed journals and proceedings in image and video coding, video processing, video streaming, and transmission. He served/serves as an associate editor for the *IEEE Transactions on Circuits and Systems for Video Technology* and *Journal of Visual Communication and Image Representation*.

Point Cloud Processing Methods for 3D Point Cloud Detection Tasks

WANG Chongchong[1], LI Yao[2], WANG Beibei[3], CAO Hong[3], ZHANG Yanyong[2]

(1. Anhui University, Hefei 230601, China;
2. University of Science and Technology of China, Hefei 230026, China;
3. Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei 230026, China)

DOI: 10.12142/ZTECOM.202304005

http://kns.cnki.net/kcms/detail/34.1294.TN.20231123.0900.002.html, published online November 23, 2023

Manuscript received: 2023-10-07

Abstract: Light detection and ranging (LiDAR) sensors play a vital role in acquiring 3D point cloud data and extracting valuable information about objects for tasks such as autonomous driving, robotics, and virtual reality (VR). However, the sparse and disordered nature of the 3D point cloud poses significant challenges to feature extraction. Overcoming limitations is critical for 3D point cloud processing. 3D point cloud object detection is a very challenging and crucial task, in which point cloud processing and feature extraction methods play a crucial role and have a significant impact on subsequent object detection performance. In this overview of outstanding work in object detection from the 3D point cloud, we specifically focus on summarizing methods employed in 3D point cloud processing. We introduce the way point clouds are processed in classical 3D object detection algorithms, and their improvements to solve the problems existing in point cloud processing. Different voxelization methods and point cloud sampling strategies will influence the extracted features, thereby impacting the final detection performance.

Keywords: point cloud processing; 3D object detection; point cloud voxelization; bird's eye view; deep learning

Citation (Format 1): WANG C C, LI Y, WANG B B, et al. Point cloud processing methods for 3D point cloud detection tasks [J]. ZTE Communications, 2023, 21(4): 38–46. DOI: 10.12142/ZTECOM.202304005

Citation (Format 2): C. C. Wang, Y. Li, B. B. Wang, et al., "Point cloud processing methods for 3D point cloud detection tasks," *ZTE Communications*, vol. 21, no. 4, pp. 38–46, Dec. 2023. doi: 10.12142/ZTECOM.202304005.

1 Introduction

3D object detection is critical for applications such as autonomous driving, robotic system navigation, and automation systems. The goal of 3D object detection is to locate and identify objects in 3D scenes. Specifically, its purpose is to estimate oriented 3D bounding boxes and semantic categories of objects from point cloud data and provide important information for subsequent analysis and processing. 3D point cloud object detection is a challenging task. Here are some major difficulties:

1) Occlusion: In complex scenes, target objects may be occluded by other objects, which affects the performance of detection algorithms.

2) Sparsity: Due to the working principle of light detection and ranging (LiDAR), point cloud data are usually sparse, which means that there are fewer effective points on the target object.

3) Point cloud noise: Noise points may be generated during the LiDAR scanning process, which will interfere with the performance of the detection algorithm.

4) Real-time requirements: 3D point cloud object detection usually needs to be completed in real time.

The object detection algorithm is a computer vision technology that can identify and locate specific objects in images or point clouds and is divided into 2D object detection[1–3] and 3D object detection[4–8]. These algorithms typically use deep learning techniques. Object detection includes tasks such as classification, localization, detection, and segmentation. Classification refers to obtaining what type of object is included in the image or point cloud data. Localization refers to the position of the given object. Detection refers to locating the position of an object and judging the category of the object. Segmentation refers to determining which object or scene each point or each pixel belongs to. Object detection algorithms are widely used in many fields, such as face recognition, automatic driving, and industrial inspection. For example, in face recognition, object detection algorithms can be used to automatically detect and track human faces and recognize the detected faces. Unmanned driving applications rely on object detection algorithms to give the poses of other traffic participants to deal with complex road conditions.

The point cloud processing method is a primary part of the

3D point cloud object detection algorithm. It can be roughly divided into the following two categories: the voxel-based point cloud processing method and point-based point cloud processing method.

The voxel processing method converts point cloud data into voxel representations, which are then processed using 3D convolutional neural networks (CNN). The point-based method is directly applied to the raw point cloud data, without converting it into grids. The point-based method can preserve the original characteristics and information of the point cloud and have lower computational cost and memory consumption.

However, the point-based method also faces some difficulties, such as dealing with the irregular structure and varying density of the point cloud and designing suitable algorithms or models for the points. Two major types of methods exist for processing the points: clustering-based methods[9 – 10] and deep learning-based methods[11 – 12]. Based on the clustering method, the appropriate clustering algorithm is selected and the clustering parameters are determined by the characteristics of the data. After denoising and merging adjacent clustering operations, the processing results are obtained. The method based on deep learning needs to prepare labeled data, select and train an appropriate deep learning model, and use the trained model to process the point cloud.

Our contribution can be summarized as follows:

1) We summarize voxel-based point cloud processing methods and find that the voxel-based methods can improve the processing performance of point clouds by optimizing the voxel partitioning scheme, improving the network structure for voxelized point clouds and the data structure for point clouds.

2) We summarize point-based point cloud processing methods and find that the point-based methods can improve the processing performance of point clouds by improving the sampling strategy of point clouds, combining the advantages of voxel-based methods, and optimizing the representation of points.

2 Basic Concepts and Metrics

1) Voxelization

Point cloud voxelization in Fig. 1 refers to the process of converting point cloud data into a voxel representation. Voxelization is to divide the point cloud into a spatially uniform voxel grid and generate many-to-one mapping between 3D points and their corresponding voxels.

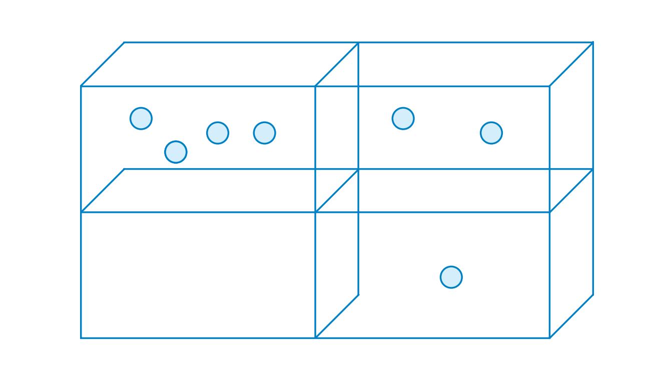
The voxelized point cloud data will be stored in memory in an orderly manner, which is beneficial to reduce random memory access and increase the efficiency of data calculation. Moreover, voxelization enables the ordered storage and downsampling of data, which allows such methods to handle much larger point cloud data. The voxelized data can also leverage spatial convolution effectively, which facilitates the extraction of local features at multiple scales and levels.

2) BEV

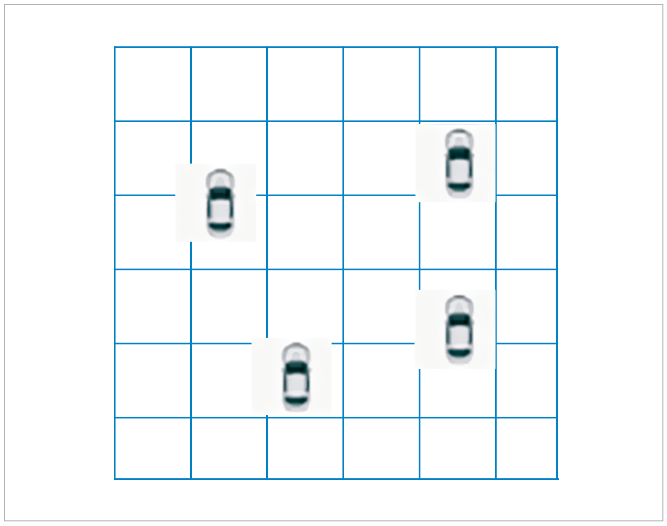
The bird's eye view (BEV) based algorithm is an advanced computer vision technique used in the field of autonomous driving. Using a combination of sensors and cameras, the algorithm creates a high-resolution overhead view of the vehicle's surroundings. The BEV perspective is shown in Fig. 2.

One of the advantages of BEV is that it provides a comprehensive view of the environment, providing a complete picture of the environment, unlike other computer vision techniques that only focus on specific objects in the environment. The perspective can provide more information for subsequent planning decisions. Another advantage is accuracy. A high-resolution top view can provide more accurate information. One disadvantage of the BEV-based algorithm is that BEV requires high computing power, which is challenging in real-time systems.

BEV is currently a very popular point cloud processing perspective. The methods related to BEV are proposed in Refs. [13 – 17]. Ref. [18] demonstrates the robustness capability of



▲Figure 1. Schematic representation of point cloud voxelization. Due to the sparsity and uneven distribution of point clouds, the number of point clouds in each voxel is unevenly distributed. There are even many voxels without point clouds. The voxel feature encoding (VFE) layer balances this through sampling



▲Figure 2. Bird's eye view based representation the BEV method.

3) FPS

Farthest point sampling (FPS) is a commonly used sampling algorithm, especially suitable for LiDAR 3D point cloud data. It can guarantee uniform sampling of samples, so it is widely used. For example, in PointNet++[12], a 3D point cloud deep learning framework, sample points are sampled by FPS and then clustered as the receptive field; in VoteNet[19], the scattered points obtained by voting are sampled by FPS and then clustered; in PVN3D[20], a 6D pose estimation algorithm, it is used to select eight feature points of the object to vote and calculate the pose.

The principle of the FPS algorithm is: Given a point cloud with N points, a point P_0 is selected from the point cloud as the starting point to obtain a sampling point set $S = \{P_0\}$. Then we calculate the distance from all points to P_0 to form an N-dimensional array L, select the point corresponding to the maximum value as P_1, and update the sampling point set $S = \{P_0, P_1\}$. Then we calculate the distance from all points to P_1. For each point P_i, if the distance to P_1 is less than $L[i]$, $L[i] = \mathrm{d}(P_i, P_1)$ is updated. Therefore, the stored L in the array is always the shortest distance from each point to the sampling point set S. The point corresponding to the maximum value in L is then selected as P_2 and the sampling point set $S = \{P_0, P_1, P_2\}$ is updated. The above steps are repeated until N' target sampling points are sampled.

Several evaluation metrics are commonly used to assess the performance of an algorithm in 3D object detection. Here are some of the most common ones:

- Average precision (AP): This is a widely used metric that measures the accuracy of object detection algorithms. It is calculated by computing the area under the precision-recall curve. AP is often used to compare the performance of different algorithms on a given dataset.
- Intersection over union (IoU): This metric measures the overlap between the predicted bounding box and the ground truth bounding box. It is calculated as the ratio of the intersection area to the union area of the two boxes. IoU is often used as a threshold to determine whether a detection is true positive or false positive.
- Mean average precision (mAP): This metric is similar to AP, but it is calculated by taking the average of AP values across multiple object categories; mAP is often used to evaluate the overall performance of an object detection algorithm.
- Precision: This metric measures the proportion of true positives among all detections. It is calculated as TP/(TP + FP), where TP is the number of true positives and FP is the number of false positives.
- Recall: This metric measures the proportion of true positives among all ground truth objects. It is calculated as TP/(TP+FN), where TP is the number of true positives and FN is the number of false negatives.

These metrics are important for evaluating 3D object detection because they provide a quantitative measure of the object performance. By comparing these metrics across different algorithms, researchers can identify which ones are most effective for a given task.

3 Voxel-Based Point Cloud Processing Methods

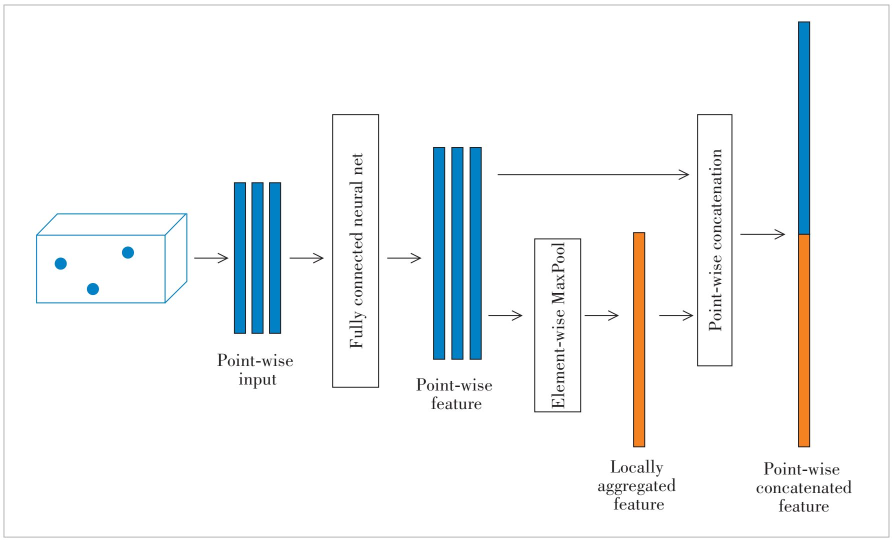
Voxel-based 3D point cloud object detection methods convert irregular point clouds into compact-shaped voxelized representations and then efficiently extract point cloud features for 3D object detection through 3D convolutional neural networks. During voxelization, the point cloud data are divided into a certain number of voxels, and these voxels are grouped and down-sampled. Since the point cloud data need to be down-sampled during the voxelization process, some detailed information will be lost. The degree of information loss is closely related to the chosen resolution.

Although the voxelization process causes information loss, it has many advantages. First, the voxelized point cloud data will be stored in an orderly manner in memory, which will help reduce random memory access and increase data computing efficiency. Second, thanks to the ordered storage and down-sampling of data brought about by voxelization, this type of method can handle point cloud data in a large amount. In addition, the voxelized data can efficiently be processed by spatial convolution, which is beneficial for extracting multi-scale and multi-level local feature information.

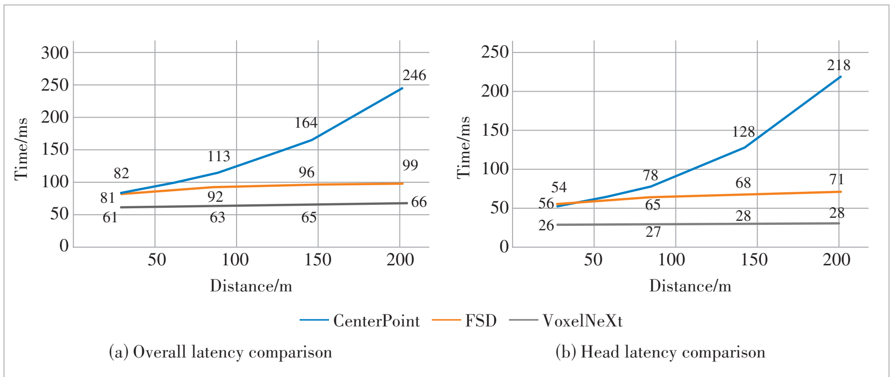
When it comes to voxel-based methods[5, 21 – 24], VoxelNet[21] has to be mentioned, which is a pioneering work. VoxelNet proposes a voxel feature encoding (VFE) layer, which groups points within a voxel in Fig. 1, and the number of point clouds after grouping is not exactly the same. In order to reduce the imbalance of the number of point clouds between groups, reduce the sampling deviation, and save computing resources, the grouped point clouds are randomly sampled so that the number of points in each group does not exceed a fixed value T. In each group, they apply PointNet[11] to learn features on each point and aggregate point features to obtain voxel-level features.

VFE is an important module. The VFE layer in Fig. 3 voxelizes the original 3D point cloud data and learns voxel-level features. This method combines the original point cloud representation and 3D voxel representation. After extracting features from the point cloud, VoxelNet uses convolutional middle layers and region proposal networks (RPNs)[25] to generate the final 3D detection box.

The key innovation of VoxelNeXt[5] is to omit the steps of anchor, sparse-to-dense, RPN, non max suppression (NMS), etc., and directly predict objects from sparse voxel features. Based on VoxelNet, VoxelNeXt has better accuracy and a speed trade-off than other detectors in nuScenes[26]. Compared with the CenterPoint[27], fully sparse 3D object detector (FSD)[28] and other methods, VoxelNeXt is more friendly to long-distance object detection in Fig. 4.


Point-wise input
Fully connected neural net
Point-wise feature
Element-wise MaxPool
Locally aggregated feature
Point-wise concatenation
Point-wise concatenated feature

▲Figure 3. Each sampled voxel (the number of point clouds $t < T$) is transformed into a feature space point by point through a fully connected neural network and then the information is aggregated from the point features to encode the surface shape contained in the voxel. The aggregated features are obtained element by element through max pooling. The point-wise feature and locally aggregated feature connection are then aggregated to get a point-wise concatenated feature



FSD: fully sparse 3D object detector

▲Figure 4. Latency on Argoverse2 and various perception ranges

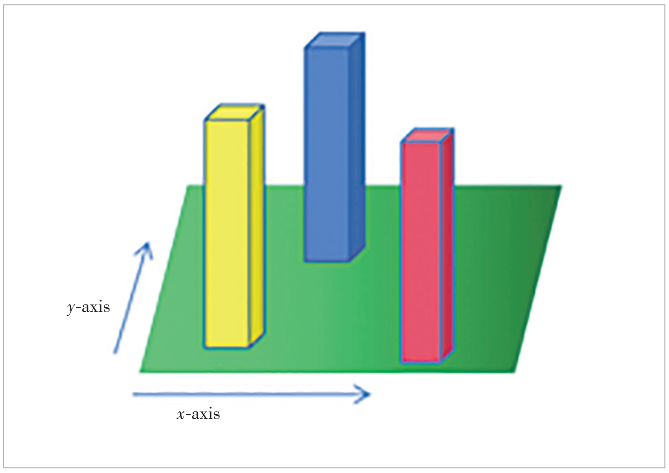
VoxelNeXt shows that fully sparse voxel-based representations are very effective for LiDAR 3D detection and tracking. VoxelNeXt proposed a fully sparse voxel-based network, which uses ordinary sparse convolutional networks for direct prediction. It uses only one extra down-sampling layer to optimize the sparse backbone network, and this simple modification enlarges the receptive field. This simple sparse linkage requires no additional parameterization layers and has a little additional computational cost.

VoxelNeXt places sparse features directly on the BEV plane and then combines features at the same location. It takes no more than 1 ms, but the effect is better than 3D sparse features. VoxelNeXt is entirely voxel-based and continuously clips irrelevant voxels along the down-sampling layer, which further saves computational resources and does not affect detection performance. Using the above-mentioned method to process voxels reduces calculation consumption without degrading performance.

The way of voxelization is not set in stone. For example, the classic VoxelNet[21] and sparsely embedded convolutional detection (SECOND)[22] divide the point cloud into a voxel to form a regular and dense voxel set, while SECOND uses sparse embedded convolution to improve efficiency.

To make a trade-off between accuracy and computation efficiency, PointPillars converts point clouds into pillars. Specifically, PointPillars divides the x axis and y axis of point cloud data into grids, and the data in the same grid is considered as a pillar (Fig. 5). This voxel division method can be considered to divide only one voxel on the z axis; P non-empty columns are generated after division; each column contains N point cloud data (more than N points are sampled as N points, and less than N points are filled with 0), and each point extracts D-dimensional features. There are nine features in PointPillar, which are $(x, y, z, r, x_c, y_c, x_p, y_p)$, where x, y and z are the 3D coordinates of the point, r is the reflection intensity, x_c and y_c are the distances from the center of the point cloud in the pillar, and x_p and y_p are the offset from the geometric center of the pillar.

In addition to improving the way of voxelization, the use of special data structures can also enhance the detection performance. The Octree-Based Transformer



▲Figure 5. Pillar division scheme of PointPillars

(OcTr) [29] algorithm first performs self-attention on the top level, constructs a dynamic octree on the hierarchical pyramid, and recursively propagates to the lower layer constrained by octants. This method can not only capture rich features from coarse-grained to fine-grained, but also control the computational complexity. Extensive experiments are conducted on Waymo Open Dataset[30] and Karlsruhe Institute of Technology and Toyota Technological Institute (KITTI) Dataset[31], and OcTr achieves new state-of-the-art results.

The performance of different detection models in three categories (Car, Pedestrian, and Cyclist) on the KITTI dataset[31] is listed in Table 1. The voxel-based method can achieve excellent performance by improving the processing method after extracting voxel features. Compared with VoxelNet, SECOND is modified to a sparse embedded convolution method to obtain a performance improvement. PointPillars proposes a way to balance accuracy and computational efficiency. A more appropriate voxel division method can achieve better results. The improvement made by OcTr is to use the transformer and the special octree data structure, which achieves better results. VoxelNeXt directly predicts objects based on sparse voxel features, without the need for sparse-to-dense conversion operations. In summary, voxelization is a method that can process large-scale point cloud data quickly and efficiently. The idea of voxelization is to process the unstructured point cloud into structured data and use the characteristics of CNN to process structured data to extract features from the point cloud. But nothing is perfect. Detailed information may be lost during the voxelization process, and voxelization is a computationally expensive step.

4 Point-Based Point Cloud Processing Methods

The point-based 3D point cloud object detection method is a method to perform object detection on the raw point cloud data. This approach preserves the unstructured form of the point cloud, but achieves a more compact representation by sampling the point cloud from its original size to smaller fixed-size N points. Sampling methods usually include random sampling and FPS, as well as several innovative sampling point methods[32].

Random sampling is achieved by randomly drawing points until N points are selected. But random sampling suffers from the scenario where points in denser regions of the point cloud are sampled more frequently than points in sparser regions of the point cloud. The FPS algorithm can mitigate this bias by using an iterative process to select points based on the furthest distance criterion. In each iteration, FPS first calculates the minimum distance from the unsampled point to the point set (the first point is randomly sampled and the second point is the point furthest from the first point) and then selects the furthest unsampled point .The final result is a more representative point cloud, but this method also suffers from expensive calculation costs.

The effect of PointNet[11] in point cloud-based methods is similar to that of VoxelNet in voxel-based methods. PointNet is a neural network-based approach that directly processes point cloud data for classification and segmentation. Operating directly on the raw point cloud eliminates unnecessary transformations of the data representation.

PointNet is a simple yet efficient point cloud feature extractor. It has three key modules: the symmetry function for unordered input, local and global information aggregation, and alignment network. Key to PointNet is that it can process unsorted point cloud data, when the disorder of point cloud is challenging in point cloud processing.

Based on Pointnet, Pointnet++[12] provides a hierarchical point cloud processing method that can effectively learn the local structure in the point cloud. Pointnet++ can handle more complex tasks such as scene segmentation, shape part segmentation, and 3D object detection.

The key technology of Pointnet++ is the introduction of hierarchical processing. Pointnet++ adopts a layered architecture. The entire point cloud is first sampled and then subdivided into smaller local areas and local features are learned on these local areas (set abstraction). Finally, these local features are aggregated to obtain global features. A Pointnet++ network consists of an encoder and a decoder. The encoder contains a collection abstraction module and the decoder contains a feature propagation module.

These two methods have promoted the application of point cloud data in the field of 3D vision and achieved remarkable research progress. There are also some extended methods[33 – 35]. Based on the PointNet series network, the feature extraction is directly applied to the original point cloud data.

▼Table 1. Performance of VoxelNet, SECOND, PointPillars and OcTr on the KITTI dataset[31]

Method	Modality	AP_{Car}			$AP_{Pedestrian}$			$AP_{Cyclist}$		
		Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
VoxelNet[21]	LiDAR	81.97	65.46	62.85	57.86	53.42	48.87	67.17	47.65	45.11
SECOND[22]	LiDAR	83.13	73.66	66.20	51.07	42.56	37.29	70.51	53.85	46.90
PointPillars[24]	LiDAR	79.05	74.99	68.30	52.08	43.53	41.49	75.78	59.07	52.92
OcTr[29]	LiDAR	88.43	78.57	77.16	61.49	57.17	52.35	85.29	70.44	66.17

AP: Average precision
KITTI: Karlsruhe Institute of Technology and Toyota Technological Institute
LiDAR: light detection and ranging
OcTr: Octree-Based Transformer
SECOND: sparsely embedded convolutional detection

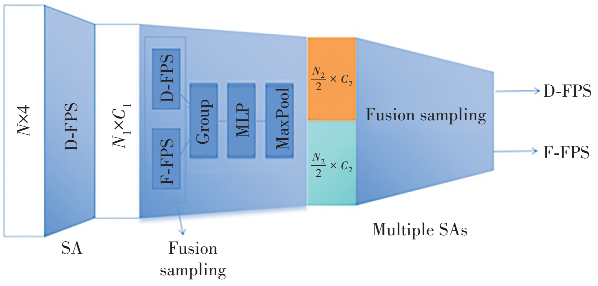
This type of method is generally divided into two steps. The first step is often to propose a rough candidate frame and the second step is to adjust and refine the position of the candidate frame.

Down-sampling operations are generally required for point cloud processing. Down-sampling can not only reduce the amount of data, but also remove some noise to improve the quality of point cloud data to a certain extent.

Commonly used point cloud down-sampling methods include random sampling, uniform sampling, furthest point sampling, etc. Different sampling methods have different advantages. Random sampling has the lowest time complexity, but retains relatively few point cloud features. Uniform sampling can preserve the overall distribution of point clouds, but the disadvantage is that it retains fewer point cloud features and cannot retain more detailed information. The advantage of furthest point sampling is that it can retain edge information. It is suitable for large-scale data processing and can quickly complete down-sampling, but it has high time complexity.

1) Improvement of the sampling method. The authors in Ref. [36] propose a lightweight and effective point-based 3D single stage object detector, named 3DSSD and believe that the feature propagation (FP) layer and refining process in the PointNet series methods[33, 37 – 38] will consume more than half of the time, but simply removing these modules and leaving only the set abstract (SA) layer to directly perform a single-stage proposal can result in a decrease in accuracy. They also believe that the down-sampling operation of the SA layer is based on the distance-based furthest point sampling method (D-FPS), which tends to retain background points. Therefore, they propose a new sampling method named F-FPS to filter background points and retain foreground points.

They use both spatial distance and semantic feature distance as the criterion in FPS. It is formulated as $C(A,B) = \lambda L_d(A,B) + L_f(A,B)$, where $L_d(A,B)$ is D-FPS, $L_f(A,B)$ is F-FPS, and λ is the balance factor. As shown in Table 2, F-FPS has the highest recall at $\lambda = 1.0$, where λ is the weight of D-FPS and F-FPS. Both the spatial distance and semantic feature distance are the criterion in FPS. In the experiment, 3DSSD adopts the method of fusion sampling. The points obtained by the two sampling methods each occupy half. The points are obtained after multi-layer SA as shown in Fig. 6. Then the candidate generation layer and two prediction heads predict the category and bounding box of the objects. 3DSSD greatly improves the speed of 3D object detection, and the speed exceeds 25 fps.



D-FPS: furthest point sampling based on 3D Euclidean distance
F-FPS: furthest point sampling based on feature distance
MLP: multi-layer perceptron
SA: set abstract

▲ Figure 6. D-FPS is first used to down-sample the point cloud once. The point cloud is sampled, grouped, MLP and maximum pooled through the 1:1 combination of D-FPS and F-FPS sampling methods. The point cloud can be sampled multiple times in the same way

▼Table 2. Points recall among different sampling strategies on the nuScenes dataset. "4 096", "1 024" and "512" represent the number of representative points in the subset. The first row of results uses only D-FPS.

Method	$\text{Recall}_{4\,096}$	$\text{Recall}_{1\,024}$	Recall_{512}
D-FPS	99.7%	65.9%	51.8%
F-FPS, λ = 0.0	99.7%	83.5%	68.4%
F-FPS, λ = 0.5	99.7%	84.9%	74.9%
F-FPS, λ = 1.0	**99.7%**	**89.2%**	**76.1%**
F-FPS, λ = 2.0	99.7%	86.3%	73.7%

D-FPS: furthest point sampling based on 3D Euclidean distance
F-FPS: furthest point sampling based on feature distance

2) Combination of point-based and voxel-based methods. There are some special methods that combine point-based and voxel-based methods[38 – 39]. Since the two methods have different advantages, their combination can bring more advantages.

The 3D object detector (STD)[38] has three main contributions. First, a spherical anchor is used to propose a point-based proposal generation example, which can achieve a high recall rate. Second, the point-based and voxel-based parts use the PointsPool link to predict efficiency and effectiveness, combining the advantages of VoxelNet[21] and PointNet[11]. Last, the alignment between classification scores and localization is achieved through a new 3D IoU prediction branch.

Point-voxel feature set abstraction for 3D object detection (PV-RCNN) [39] is a high-performance 3D object detection framework. It integrates the method of point-cloud voxelization and convolution and the method of PointNet-based set abstraction to obtain better point cloud features. PV-RCNN directly uses the original point cloud, processes the point cloud through 3D sparse convolution after voxelization and performs classification and box prediction through RPN on the BEV plane. At the same time, the FPS is used for key point sampling, and the key point features and the features of non-empty voxels around the key points are collected through the VSA module. These features are used to make up for the information loss during voxelization. Object category and bounding box predictions are refined through a two-part combination.

Both PV-RCNN and STD have achieved good results on the KITTI dataset (Table 3), and their performance outperforms either the voxel-based or point-cloud-based method used alone,

▼**Table 3. Performance testing on the KITTI test set. Mean average precision is taken as the evaluation metric. The table shows better performance of PV-RCNN and STD**

Method	$AP_{Car\text{-}3D\ Detection}$			$AP_{C\ ar\text{-}BEV\ Detection}$			$AP_{Cyclist\text{-}3D\ Detection}$			$AP_{Cyclist\text{-}BEV\ Detection}$		
	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
SECOND[22]	83.34	72.55	65.82	89.39	83.77	78.59	71.33	52.08	45.83	76.50	56.05	49.45
Fast Point R-CNN[35]	85.29	77.40	70.24	90.87	87.84	80.52	-	-	-	-	-	-
STD[38]	87.95	79.71	75.09	94.74	89.19	**86.42**	78.69	61.59	55.30	81.36	67.23	59.35
PV-RCNN[39]	**90.25**	**81.43**	**76.82**	**94.98**	**90.65**	86.14	78.60	**63.71**	57.65	82.49	**68.89**	**62.41**

AP: average precision
BEV: bird's eye view
PV-RCNN: point-voxel feature set abstraction for 3D object detection
R-CNN: Region-CNN
SECOND: Sparsely Embedded Convolutional Detection
STD: Sparse-to-Dense 3D Object Detector for Point Cloud

demonstrating the benefits of combining these two complementary methods.

Besides STD and PV-RCNN, the methods proposed in Refs. [40] and [41] also combine the point-based and voxel-based manners.

Some methods use other networks such as graph neural networks (GNNs). They convert the point cloud into a regular grid or voxel and use CNN (point cloud representation in the grid) or deep learning technology to process the point cloud (point cloud in the point set) after obtaining the point set through sampling and other operations (Fig. 7). In addition, Point-GNN[42] constructs the point cloud into a graph. It has three main components: graph construction from point cloud, graph neural network for object detection, and bounding box merging and scoring. Specifically, the points in the point cloud are used as N vertices, with a point as the center and r as the radius. Neighboring points in the range are concatenated to construct a graph $G = (P, E)$, for example:

$$E = \left\{ \left(p_i, p_j \right) \middle| \left| x_i - x_j \right|^2 < r \right\}.$$

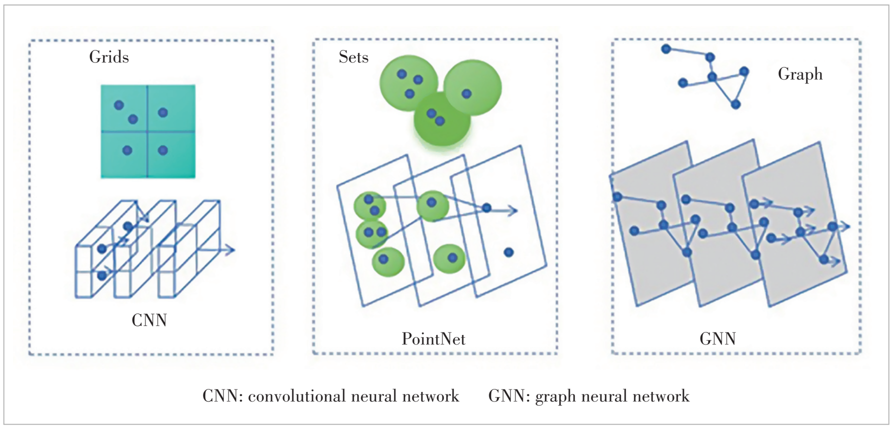
In order to reduce complexity, Point-GNN uses voxel operations to down-sample point clouds in the actual process and the voxels are only used for reducing the point cloud density. Once constructed, the point cloud is processed using a multi-iteration GNN[43].

Point-GNN has achieved excellent performance on the KITTI test data set. The average precision of the car, pedestrian and cyclist at the easy level reached 88.33, 51.92 and 78.60, respectively, at the modality levels 79.47, 43.77 and 63.48, respectively, and at the hard level 72.29, 40.14 and 57.08, respectively. The detection performance of the car and cyclist surpasses both the radar-only methods such as STD[38] and PointRCNN[33] and the radar and image fusion methods such as AVOD-FPN[44] and UberATG-MMF[45].

The point-based methods still have several modules that need to be improved. One module is sampling, which can reduce the consumption of computing resources by selecting a subset of points from the original point cloud. However, sampling may cause some information loss, which affects the quality of the features that can be extracted in subsequent operations. Therefore, the choice of the sampling algorithm is crucial for the point-based method. For example, PointNet++ uses feature propagation to suppress the information loss caused by sampling, and 3DSSD improves different sampling methods to retain more useful information and improve efficiency.

Another module that can be improved is voxelization, which is a special method to introduce voxels into point cloud processing. Voxels are small cubes that divide the three-dimensional space and contain a certain number of points. The advantage of voxelization is to convert point clouds into ordered data and also reduce computational complexity. However, voxelization may introduce quantization errors and lose some fine-grained details. Therefore, some methods combine the information obtained from both voxels and points to improve performance. For example, PV-RCNN uses voxel-based RPN and point-based RoI feature extractors (RoIFEs) to achieve state-of-the-art results on 3D object detection.

The third module that can be improved



CNN: convolutional neural network GNN: graph neural network

▲**Figure 7. Representation of point-cloud grids, sets and graph and their corresponding processing methods** is the basic network model, which is used to process the point cloud data and extract features. Different network models have different advantages and disadvantages for point cloud processing. For example, GNN can capture the structure and relationship of point cloud data by using nodes and edges. It can handle irregular and unordered data better than convolutional neural networks.

5 Conclusions

In this paper, we summarize the processing of point clouds in object detection. Point cloud processing is the first step in most models and it can greatly affect the performance of subsequent detection operations. Point cloud processing can be divided into two categories: voxel-based and point-based processing, both of which have their own advantages and disadvantages.

Voxel-based processing is a method that divides the three-dimensional space into small cubes called voxels and assigns points to voxels according to their coordinates. The advantage of voxel-based processing is that it can convert point clouds into ordered data and reduce computational complexity. However, voxel-based processing may introduce quantization errors and lose some fine-grained details. Many works have improved voxel-based methods by changing the way voxels are divided, changing the network for processing voxels, changing the data structure for processing data, etc. These approaches can reduce time complexity and organize voxel-level features well, further improving performance.

Point-based processing is a method that directly operates on raw points without any transformation or quantization. The advantage of point-based processing is that it can preserve the original structure and information of point clouds. However, point-based processing may face challenges such as irregularity and sparsity of point clouds. Many works have improved point-based methods by improving the way of point cloud sampling, introducing some voxel-based features or directly obtaining the graph structure from the structure of the original point cloud data. These approaches can enhance the feature extraction and representation of points, which can also significantly improve the performance of subsequent detection.

References

[1] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks [J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(6): 1137 – 1149. DOI: 10.1109/tpami.2016.2577031

[2] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection [EB/OL]. (2016-05-09)[2023-08-20]. https://arxiv.org/abs/1506.02640

[3] LAW H, DENG J. CornerNet: detecting objects as paired keypoints [J]. International journal of computer vision, 2020, 128(3): 642 – 656. DOI: 10.1007/s11263-019-01204-1

[4] SHI S S, WANG X G, LI H S. PointRCNN: 3D object proposal generation and detection from point cloud [EB/OL]. (2019-05-16)[2023-08-21]. https://arxiv.org/abs/1812.04244

[5] CHEN Y K, LIU J H, ZHANG X Y, et al. VoxelNeXt: fully sparse VoxelNet for 3D object detection and tracking [EB/OL]. (203-03-20)[2023-08-21]. https://arxiv.org/abs/2303.11301

[6] YANG Z T, SUN Y N, LIU S, et al. STD: sparse-to-dense 3D object detector for point cloud [EB/OL]. (2019-07-22) [2023-08-21]. https://arxiv. org/abs/1907.10471

[7] PHILION J, FIDLER S. Lift, splat, shoot: encoding images from arbitrary camera rigs by implicitly unprojecting to 3D [C]//European Conference on Computer Vision. Springer, 2020: 194 – 210.10.1007/978-3-030-58568-6_12

[8] YIN T W, ZHOU X Y, KRÄHENBÜHL P. Center-based 3D object detection and tracking [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2021: 11779 – 11788. DOI: 10.1109/CVPR46437.2021.01161

[9] KLASING K, WOLLHERR D, BUSS M. A clustering method for efficient segmentation of 3D laser data [C]//2008 IEEE International Conference on Robotics and Automation. IEEE, 2008: 4043 – 4048. DOI: 10.1109/ROBOT.2008.4543832

[10] KLASING K, WOLLHERR D, BUSS M. Realtime segmentation of range data using continuous nearest neighbors [C]//2009 IEEE International Conference on Robotics and Automation. IEEE, 2009: 2431 – 2436. DOI: 10.1109/ROBOT.2009.5152498

[11] CHARLES R Q, HAO S, MO K C, et al. PointNet: deep learning on point sets for 3D classification and segmentation [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017: 77 – 85. DOI: 10.1109/CVPR.2017.16

[12] QI C R, YI L, SU H, et al. PointNet++: deep hierarchical feature learning on point sets in a metric space [EB/OL]. (2017-06-07)[2020-08-21]. https://arxiv.org/abs/1706.02413

[13] HUANG J J, HUANG G, ZHU Z, et al. BEVDet: high-performance multi-camera 3D object detection in bird-eye-view [EB/OL]. (2022-06-16)[2023-08-21]. https://arxiv.org/abs/2112.11790

[14] LI Z Q, WANG W H, LI H Y, et al. Bevformer: learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers [EB/OL]. (2022-07-13)[2023-08-21]. https://arxiv.org/abs/2203.17270

[15] LI Y H, GE Z, YU G Y, et al. BEVDepth: acquisition of reliable depth for multi-view 3D object detection [EB/OL]. (2022-11-30)[2023-08-21]. https://arxiv.org/abs/2206.10092

[16] LIU Z J, TANG H T, AMINI A, et al. BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation [EB/OL]. (2022-06-16)[2023-08-21]. https://arxiv.org/abs/2205.13542

[17] WANG R H, QIN J, LI K Y, et al. BEV-LaneDet: a simple and effective 3D lane detection baseline [EB/OL]. (203-03-11)[2023-08-21]. https://arxiv. org/abs/2210.06006

[18] DONG Y P, KANG C X, ZHANG J L. Benchmarking robustness of 3D object detection to common corruptions in autonomous driving [EB/OL]. (2023-03-20)[2023-08-21]. https://arxiv.org/abs/2303.11040

[19] QI C R, LITANY O, HE K M, et al. Deep hough voting for 3D object detection in point clouds [EB/OL]. (2019-08-22) [2023-08-21]. https://arxiv. org/abs/1904.09664

[20] HE Y S, SUN W, HUANG H B, et al. PVN3D: deep point-wisea 3D keypoints voting network for 6DoF pose estimation [EB/OL]. (2020-03-24)[2023-08-21]. https://arxiv.org/abs/1911.04231

[21] ZHOU Y, TUZEL O. VoxelNet: end-to-end learning for point cloud based 3D object detection [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2018: 4490 – 4499. DOI: 10.1109/CVPR.2018.00472

[22] YAN Y, MAO Y X, LI B. SECOND: sparsely embedded convolutional detection [J]. Sensors, 2018, 18(10): 3337. DOI: 10.3390/s18103337

[23] SIMON M, AMENDE K, KRAUS A, et al. Complexer-YOLO: real-time 3D object detection and tracking on semantic point clouds [EB/OL]. (2019-04-16)[2023-08-21]. https://arxiv.org/abs/1904.07537

[24] LANG A H, VORA S, CAESAR H, et al. PointPillars: fast encoders for object detection from point clouds [EB/OL]. (2020-03-24)[2023-08-21]. https://arxiv.org/abs/1812.05784

[25] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time ob-

ject detection with region proposal networks [J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(6): 1137 – 1149. DOI: 10.1109/TPAMI.2016.2577031

[26] CAESAR H, BANKITI V, LANG A H, et al. nuScenes: a multimodal dataset for autonomous driving [EB/OL]. (2020-05-05)[2023-08-21]. https://arxiv.org/abs/1903.11027

[27] YIN T W, ZHOU X Y, KRÄHENBÜHL P. Center-based 3D object detection and tracking [EB/OL]. (2021-01-06) [2023-08-21]. https://arxiv. org/abs/2006.11275

[28] FAN L, WANG F, WANG N Y, et al. Fully sparse 3D object detection [EB/OL]. (2022-10-03)[2023-08-21]. https://arxiv.org/abs/2207.10035

[29] ZHOU C, ZHANG Y N, CHEN J X, et al. OcTr: octree-based transformer for 3D object detection [EB/OL]. (2023-03-22)[2023-08-21]. https://arxiv.org/pdf/2303.12621.pdf

[30] SUN P, KRETZSCHMAR H, DOTIWALLA X, et al. Scalability in perception for autonomous driving: waymo open dataset [EB/OL]. (2023-03-22)[2023-08-21]. https://arxiv.org/abs/1912.04838

[31] GEIGER A, LENZ P, URTASUN R, et al. Are we ready for autonomous driving? The KITTI vision benchmark suite [C]//2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012. DOI: 10.1109/CVPR.2012.6248074

[32] YANG Z T, SUN Y N, LIU S, et al. 3DSSD: point-based 3D single stage object detector [EB/OL]. (2020-02-24)[2023-08-21]. https://arxiv.org/abs/2002.10187

[33] SHI S S, WANG X G, LI H S. PointRCNN: 3D object proposal generation and detection from point cloud [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020: 770 – 779. DOI: 10.1109/CVPR.2019.00086

[34] QI C R, LIU W, WU C X, et al. Frustum PointNets for 3D object detection from RGB-D data [EB/OL]. (2018-04-13)[2023-08-21]. https://arxiv. org/abs/1711.08488

[35] CHEN Y L, LIU S, SHEN X Y, et al. Fast point R-CNN [EB/OL]. (2019-08-16)[2023-08-21]. https://arxiv.org/abs/1908.02990

[36] YANG Z T, SUN Y N, LIU S, et al. 3DSSD: point-based 3D single stage object detector [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020: 11037 – 11045. DOI: 10.1109/CVPR42600.2020.01105

[37] YANG Z T, SUN Y N, LIU S, et al. IPOD: intensive point-based object detector for point cloud [EB/OL]. (2018-12-13)[2023-08-21]. https://arxiv. org/abs/1812.05276

[38] YANG Z T, SUN Y N, LIU S, et al. STD: sparse-to-dense 3D object detector for point cloud [C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2020: 1951 – 1960. DOI: 10.1109/ICCV.2019.00204

[39] SHI S S, GUO C X, JIANG L, et al. PV-RCNN: point-voxel feature set abstraction for 3D object detection [EB/OL]. (2021-04-09)[2023-08-21]. https://arxiv. org/abs/1912.13192

[40] YOU Y R, WANG Y, CHAO W-L, et al. Pseudo-lidar++: accurate depth for 3d object detection in autonomous driving [EB/OL]. (2020-02-15)[2023-08-21]. https://arxiv.org/abs/1906.06310

[41] HE C H, ZENG H, HUANG J Q, et al. Structure aware single-stage 3D object detection from point cloud [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020: 11870 – 11879. DOI: 10.1109/CVPR42600.2020.01189

[42] SHI W J. Point-GNN: graph neural network for 3D object detection in a point cloud [EB/OL]. (2020-03-02)[2023-08-21]. https://arxiv.org/abs/2003.01251

[43] SCARSELLI F, GORI M, TSOI A C, et al. The graph neural network model [J]. IEEE transactions on neural networks, 2009, 20(1): 61 – 80. DOI: 10.1109/tnn.2008.2005605

[44] KU J, MOZIFIAN M, LEE J, et al. Joint 3D proposal generation and object detection from view aggregation [C]//2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). ACM, 2018: 1 – 8. DOI: 10.1109/IROS.2018.8594049

[45] LIANG M, YANG B, CHEN Y, et al. Multi-task multi-sensor fusion for 3D object detection [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020: 7337 – 7345. DOI: 10.1109/CVPR.2019.00752

Biographies

WANG Chongchong received his BS degree in computer science and technology from Huazhong Agricultural University, China in 2022. He is currently pursuing a master’s degree in computer science and technology at Anhui University, China.

LI Yao received his BS degree in electronic information engineering from Jilin University, China in 2019. He is currently pursuing a PhD degree in computer science and technology at the University of Science and Technology of China. His research interests include computer vision and intelligent transportation perception system.

WANG Beibei graduated from University of Science and Technology of China with BS in physics in 2014. He furthered his studies at the University of Southern California, USA, where he obtained PhD in physics in 2020 and MS in computer science in parallel. His research interests currently focus on computer vision and multimodal perception methods for autonomous systems.

CAO Hong received his PhD degree from Zhejiang University, China in 2014. He is currently a associate research fellow with the Institute of Artificial Intelligence, Hefei Comprehensive National Science Center (Anhui Artificial Intelligence Laboratory), China. His research interests include autonomous driving, roadside perception and robotics.

ZHANG Yanyong (yanyongz@ustc.edu.cn) received her BS from the University of Science and Technology of China (USTC) in 1997, and PhD from Penn State University in 2002. From 2002 and 2018, she was on the faculty of the Electrical and Computer Engineering Department at Rutgers University, USA. She was also a member of the Wireless Information Networks Laboratory (Winlab). Since July 2018, she joined the school of Computer Science and Technology at USTC. She has 21 years of research experience in the areas of sensor networks, ubiquitous computing, and high-performance computing, and has published more than 140 technical papers in these fields. She received the NSF CAREER award in 2006, and was elevated to IEEE Fellow in 2017. She has served/currently serves as the Associate Editor for several journals, including *IEEE/ACM Transactions on Networking*, *IEEE Transactions on Mobile Computing*, *IEEE Transactions on Service Computing*, *IEEE Transactions on Dependable and Secure Computing*, and *Elsevier Smart Health*. She has served on various conference TPCs including DSN, Sensys, Infocom, etc. She is the TPC co-chair of IPSN’22.

Perceptual Optimization for Point-Based Point Cloud Rendering

YIN Yujie, CHEN Zhang

(School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710129, China)

DOI: 10.12142/ZTECOM.202304006

https://kns.cnki.net/kcms/detail/34.1294.TN.20231109.1549.001.html, published online November 10, 2023

Manuscript received: 2023-10-05

Abstract: Point-based rendering is a common method widely used in point cloud rendering. It realizes rendering by turning the points into the base geometry. The critical step in point-based rendering is to set an appropriate rendering radius for the base geometry, usually calculated using the average Euclidean distance of the N nearest neighboring points to the rendered point. This method effectively reduces the appearance of empty spaces between points in rendering. However, it also causes the problem that the rendering radius of outlier points far away from the central region of the point cloud sequence could be large, which impacts the perceptual quality. To solve the above problem, we propose an algorithm for point-based point cloud rendering through outlier detection to optimize the perceptual quality of rendering. The algorithm determines whether the detected points are outliers using a combination of local and global geometric features. For the detected outliers, the minimum radius is used for rendering. We examine the performance of the proposed method in terms of both objective quality and perceptual quality. The experimental results show that the peak signal-to-noise ratio (PSNR) of the point cloud sequences is improved under all geometric quantization, and the PSNR improvement ratio is more evident in dense point clouds. Specifically, the PSNR of the point cloud sequences is improved by 3.6% on average compared with the original algorithm. The proposed method significantly improves the perceptual quality of the rendered point clouds and the results of ablation studies prove the feasibility and effectiveness of the proposed method.

Keywords: point cloud rendering; outlier detection; perceptual optimization; point-based rendering; perceptual quality

Citation (Format 1): YIN Y J, CHEN Z. Perceptual optimization for point-based point cloud rendering [J]. ZTE Communications, 2023, 21(4): 47 – 53. DOI: 10.12142/ZTECOM.202304006

Citation (Format 2): Y. J. Yin, Z. Chen, "Perceptual optimization for point-based point cloud rendering," *ZTE Communications*, vol. 21, no. 4, pp. 47 – 53, Dec. 2023. doi: 10.12142/ZTECOM.202304006.

1 Introduction

With the increasing demand for stereoscopic visual effects and immersive experiences, 3D point clouds are emerging as one of the primary three-dimensional data in various areas. Point clouds are usually generated from sensor data, e. g., triangulation and Time-of-Flight (ToF)[1]. A point cloud is a set of points in a given coordinate system representing information about a scene or an object, often complemented with per-point attributes. This set of points may define the shape of real or computer-generated objects or even complex scenes[2]. A point cloud has both geometric and attribute features, where the geometric features reflect the number of points in the point cloud in relation to their position, while the attribute features can contain additional information such as the color, grey value, average vector and reflectivity of the points. 3D point clouds[3–5] have found broad applications in manufacturing, construction, environmental monitoring, navigation, animation, rendering, etc.[6] However, various distortions may be introduced during the acquisition, compression[7], transmission, storage, and rendering processes of point clouds, leading to degraded perceptual quality[6]. Improving the Quality of Experience (QoE)[8] of point clouds is very important for related applications.

Point cloud rendering is the process of producing a visual representation that can be consumed by users using an available display, e.g., conventional 2D, stereo, auto-stereoscopic, head-mounted displays, etc.[9] The rendering process has a significant impact on the quality perceived by the user[10]. Point clouds have many properties that differ from traditional video images. First, a point cloud is an unordered collection. Second, a point cloud is unstructured, and the relative positions between individual neighboring points are not the same, which means that even finding neighboring points requires traversing all the points in the point cloud. Finally, point clouds are non-uniform. They are usually obtained by sampling a single object, and the densities of points in different regions of the same point cloud may vary significantly[11]. These characteristics of point cloud data pose new requirements and higher challenges to the algorithms for point cloud rendering.

Currently, two main categories of rendering methods are commonly used for point clouds[11]. One is mesh-based rendering, which constructs a polygonal mesh through surface reconstruction algorithms, e. g., Poisson surface reconstruction[12]. These rendering methods can construct a completely closed surface, but the results highly depend on the reconstruction algorithm, and human intervention may be required to reconstruct complex surfaces. As the object in 3D may be complicated, extensive computation is generally involved[13]. The other category is point-based rendering, which achieves rendering results by turning points into the base geometry with area (e.g., circles, spheres, cubes, etc.). The point-based methods are low in computation and easy to implement, but it is necessary to define the size of the base geometry in advance. Otherwise, empty spaces will appear between points (size too small) or aliasing artefacts (size too large). Generally speaking, point-based rendering is easier to perform in comparison with a polygonal mesh representation where surface reconstruction and interpolation are usually needed[10].

Point-based rendering is a common method widely used in point cloud rendering. In Ref. [10], JAVAHERI et al. used the average Euclidean distance of N (N=10) nearest neighboring points to determine the rendering radius of points in their research work of point clouds, thus solving the problem of the requirement of artificially defining the size of the base geometry in point rendering and avoiding the empty spaces between points.

In point cloud rendering, the rendering radius is generally determined by the nearest neighbors. However, this approach faces challenges when there are outliers. Since the rendering radius is determined by the nearest-neighboring points, the rendering radius of the outliers far away from the central region of the point cloud sequence may be set as a very large value, which impacts the subjective quality. The above problem is illustrated in Fig. 1. The rendering radius of the outliers can seriously affect the subjective quality of the rendered point cloud sequence. Therefore, it is important to filter out the outliers and set an appropriate rendering radius before rendering to improve the perceptual quality of the point cloud sequence.

To solve the problem of poor perceptual quality in point-based rendering due to the large rendering radius of outlier points, we propose an algorithm for point cloud rendering through outlier detection to optimize the perceptual quality of rendering. The algorithm constructs outlier detection conditions by combining local and global geometric features. During rendering, the rendering radius of outliers is set to the minimum. By doing so, the proposed method realizes the perceptual optimization of point cloud rendering and improves the perceptual quality after point cloud rendering.

The main contributions of this paper are as follows:

1) We propose an outlier detection algorithm with low complexity. The proposed method uses global and local geometric features to detect outliers in the sequence.

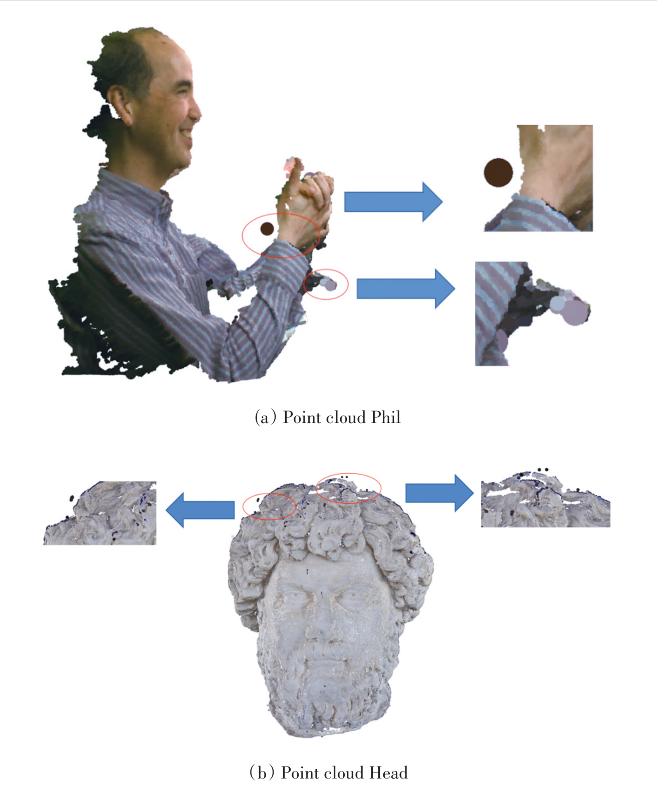
2) We apply the outlier detection algorithm to point cloud rendering to optimize the perceptual quality of the point rendering. Compared with the original rendering algorithm, our method improves both the perceptual quality of the rendered point cloud and the objective quality. The ratio of peak signal-to-noise ratio (PSNR) improvement is about 3.6%.

The rest of this paper is organized as follows. Section 2 introduces the rendering algorithm with outlier point detection. Section 3 shows the rendering results of this rendering algorithm and examines the algorithm's performance in terms of both objective and perceived quality. The paper is summarized in Section 4.

2 Rendering Based on Outlier Detection

2.1 Overview

Suppose the coordinates of the i-th point in the point cloud sequence are (x_i, y_i, z_i), $i = 1,2,\cdots,M$. The coordinates of the nearest N neighboring points (including the point itself) are $\left(x_i^n, y_i^n, z_i^n\right)$, $n = 1,2,\cdots,N$, where i stands for the point to be detected in the point cloud sequence and n stands for the n-th neighboring point of the point to be detected.


(a) Point cloud Phil

(b) Point cloud Head

▲Figure 1. Illustration of outliers in rendering (rendering primitive is a circle)

The average Euclidean distance of the N nearest neighboring points is taken as the rendering radius of the point. Thus the rendering radius r_i of the i-th point is:

$$r_i = \sqrt{\frac{1}{N} \times \sum_{n=1}^{N} \left\| \left(x_i^n, y_i^n, z_i^n\right) - \left(x_i, y_i, z_i\right) \right\|^2} . \tag{1}$$

As can be seen from the above equation, outlier points far away from the central region of the point cloud sequence will have too sizeable Euclidean distance from their nearest neighbors, and thus, the rendering radius r_i will be too large, which in turn results in an impaired subjective perception of the quality of the rendered point cloud sequence.

In order to solve the problem of too large a rendering radius of outlier points in the above method, we need to add a process of detecting and determining outlier points before rendering each point in the point cloud sequence. We determine whether each point is an outlier, filter the outlier points and change their rendering radii. For the outlier points, the minimum rendering radius is used. For the non-outlier points, the radius of the nearest neighbor rendering r_i is used as shown in Eq. (1).

Outlier detection is a commonly used detection method in data analysis and processing, often used to identify abnormal samples or points in outlier states that significantly deviate from the central region of the sequence. Many outlier mining methods can be categorized into five groups: distribution-based, depth-based, clustering-based, distance-based, and density-based[14 – 15]. Commonly used detection methods are distance detection-based K-nearest neighbor algorithms, isolated forest methods, clustering-based DBSCAN algorithms[16], LOF algorithms[17], supervised or unsupervised algorithms based on machine learning[18], and distribution-based and density-based outlier detection[19]. However, the above methods have the disadvantages of high time complexity, poor detection results in high-dimensional sparseness, and insignificant mathematical geometric features.

Since determining the rendering radius in terms of the nearest neighboring points is through geometric features, detecting outlier states at each point based on geometric features is reasonable. To ensure the accuracy of detection, global and local geometric features construct a comprehensive judgment condition to quickly check and judge whether the current point is in an outlier state.

The implementation complexity of the detection method needs to be as low as possible for rendering. So we use discrete point detection techniques based on statistical distributions, combining local and global geometric features to judge outliers. This is a method that has solid probabilistic statistical theory support. After modeling, it does not require model-based data, and only the minimum amount of information describing the model needs to be stored, which can effectively reduce the data storage.

2.2 Construction of Outlier Judgment Conditions

The basic principle of the algorithm for point cloud rendering through outlier detection in this paper is to determine whether the detected point is in an outlier state by judging whether the point satisfies the distribution type of the dataset and, thus, whether the point is in an outlier state. Assuming that the data satisfy the condition of obeying normal distribution, p_0 belongs to the dataset p, and the mean and standard deviation of this dataset are $\bar{p}$ and σ, respectively. The judgment criteria of outlier detection are shown in Eq. (2)[20]. When Eq. (2) is not satisfied, the point p_0 is determined to be in an outlier state.

$$\left| \frac{p_0 - \bar{p}}{\sigma} \right| \leqslant 3 . \tag{2}$$

From Eq. (1), we can see that the rendering method based on the nearest neighbor distance determines the size of the rendering radius of each point in the point cloud sequence by the geometric density feature in the local area. Therefore, the geometric features of outlier points are also chosen as density features in this method, i.e., the rendering radius is calculated based on the nearest-neighbor distance.

Let the rendering radius of the nearest N (N = 10) neighbors of the detection point be r_i^n, the mean value of the local geometric density feature be $E(r)$, and the standard deviation be σ_r. $E(r)$ and σ_r are calculated in Eqs. (3) and (4).

$$E(r) = \frac{1}{N} \sum_{n=1}^{N} r_i^n , \tag{3}$$

$$\sigma_r = \sqrt{\frac{\sum_{n=1}^{N} \left\| r_i^n - E(r) \right\|^2}{N}} . \tag{4}$$

Combining the outlier judgment criteria, we substitute the rendering radius data into Eq. (2), and the outlier judgment criteria for the local geometric feature construction are shown as follows[12]:

$$\left| \frac{r_i - E(r)}{\sigma_r} \right| = \left| \frac{r_i - \frac{1}{N} \sum_{n=1}^{N} r_i^n}{\sqrt{\frac{\sum_{n=1}^{N} \left\| r_i^n - E(r) \right\|^2}{N}}} \right| \leqslant 3 . \tag{5}$$

When the detection point does not satisfy the local determination condition of the above equation, it is determined that the detection point is in the outlier state.

Since the number of detected points in the point set covered by the local geometric features is small, only local features are not enough to detect outliers. In addition, the local geometric features easily fall into the local small sample situation. When the points included in the local geometric features are all outliers, the local judgment condition criteria will fall into local condition satisfaction, resulting in outlier detection failure. Therefore, it is limited to judging whether the detected points are in an outlier state only by local geometric conditions, and it is necessary to expand the geometric features and the capacity of the point set in the judgment conditions. Therefore, the global geometric features are added as a judgment condition for outlier detection.

Let the global geometric feature be R, which is calculated in Eq. (6), where M is the number of bits in the global geometric point set.

$$R = \frac{1}{M}\sum_{i=1}^{M} r_i. \tag{6}$$

In order to reduce the complexity of the algorithm implementation, the actual rendering process can use the cumulative means as a global geometric feature, which is calculated in Eq. (7).

$$R = \frac{1}{m}\sum_{i=1}^{m} r_i, \tag{7}$$

where m is the ranking bit of the current detection point in the rendering process.

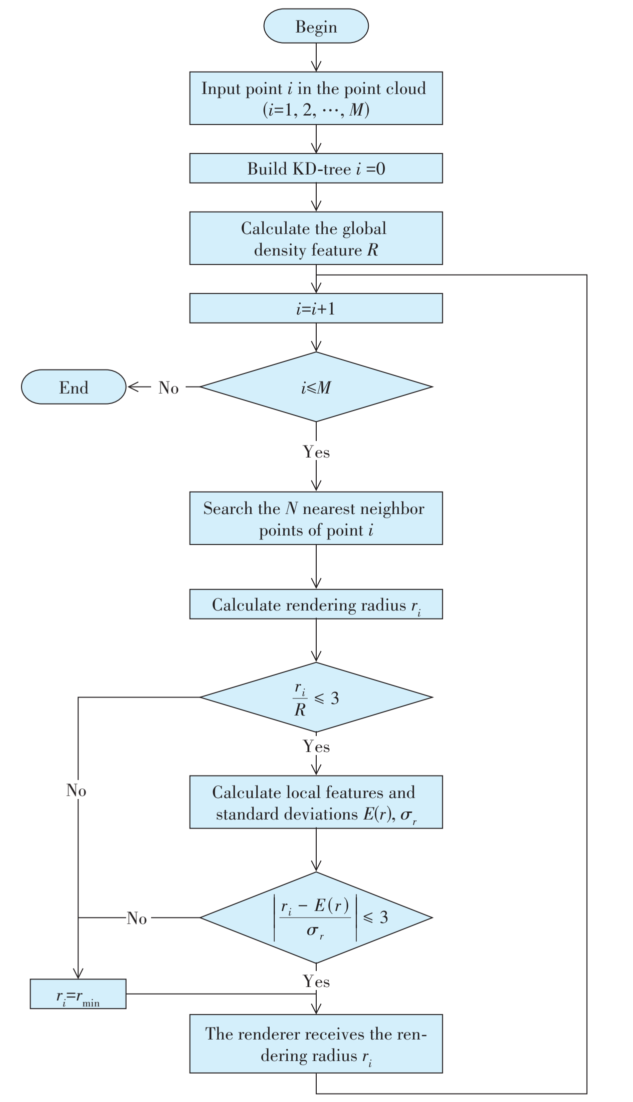
Therefore, based on Eq. (2), the criteria for determining outlier points for global geometric feature construction are:

$$\frac{r_i}{R} = \frac{r_i}{\frac{1}{m}\sum_{i=1}^{m} r_i} \leqslant 3 . \tag{8}$$

When a detection point does not satisfy any of the determination conditions of the global condition and the local condition, the detection point is judged to be an outlier.

2.3 Rendering Through Outlier Detection

Before rendering the points in the point cloud sequence, the outlier status is detected for each point. First, we input the points in the point cloud sequence. Since the point cloud is disordered, we need to arrange the disordered points in the point cloud sequence in an orderly manner to facilitate fast and efficient retrieval. Therefore, the k-dimensional (KD)-tree[21] is constructed with the detection points as the core, and the points are divided according to the tree structure to facilitate the subsequent nearest neighbor search. In constructing the KD-tree, we follow the method in Ref. [20] to transform the distribution of point sets in the point cloud sequence into a normal distribution. Thus, the prerequisite assumption of normal distribution of data in Eq. (2) is satisfied. The global geometric feature R is calculated according to Eq. (7). The points in the point cloud sequence are traversed according to the order in the KD-tree, and the rendering radius r_i is calculated according to Eq. (1) after searching the N nearest neighboring points of the current detection point. The local feature values $E(r)$ and σ_r are calculated according to Eqs. (3) and (4). The global and local conditions are judged according to Eqs. (5) and (8) in turn, and if either of the two judging conditions is not satisfied, the detected point is judged to be in the outlier state. The rendering radius r_i of the point in the outlier state is set to the minimum $r_{\min}$. When the detected point satisfies both local and global detection conditions, the rendering radius r_i is not changed. Finally, the detected point rendering radius r_i is passed to the renderer for the rendering of each point by the renderer. The specific flow of the rendering algorithm through outlier point detection is shown in Fig. 2.


Begin
Input point i in the point cloud (i=1, 2, ⋯, M)
Build KD-tree i =0
Calculate the global density feature R
i=i+1
End
No
i≤M
Yes
Search the N nearest neighbor points of point i
Calculate rendering radius r_i
$\frac{r_i}{R} \leqslant 3$
Yes
No
Calculate local features and standard deviations $E(r)$, σ_r
No
$\left|\frac{r_i - E(r)}{\sigma_r}\right| \leqslant 3$
Yes
r_i=$r_{\min}$
The renderer receives the rendering radius r_i

KD: k-dimensional

▲**Figure 2. Flow chart of proposed rendering through the outlier detection**

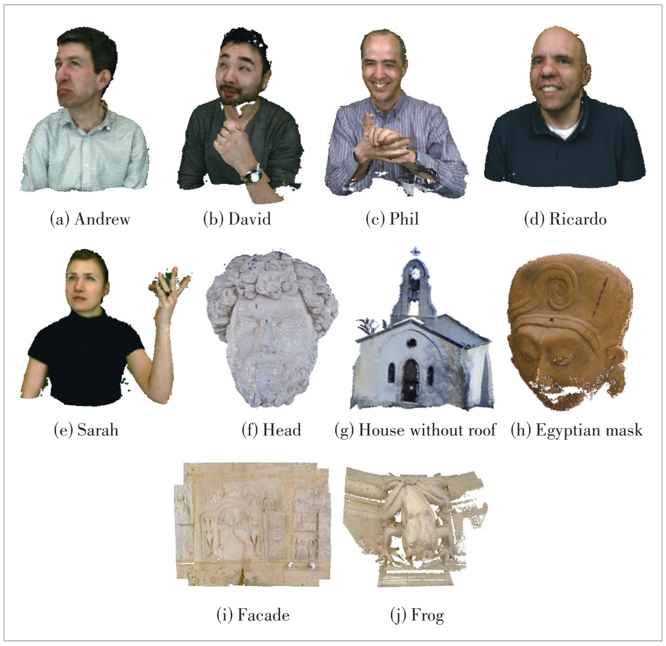
3 Experimental Results

3.1 Point Cloud Sequence Selection

We use ten static point clouds to examine the performance of the proposed method in terms of objective and perceptual quality. Fig. 3 shows a thumbnail of the point cloud sequences. Table 1 shows the detailed information of each point cloud sequence.

3.2 Subjective and Objective Results

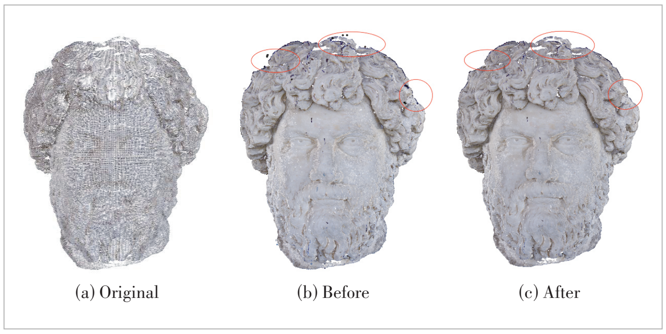
The effect of the rendering through outlier detection is first evaluated by conducting a subjective comparison before and after outlier optimization. Take Head's character point cloud sequence and Phil's object point cloud sequence as examples. Figs. 4 and 5 show the subjective comparison effects of the above two point cloud sequences using the original point-rendering[10] and the proposed method. It can be found that the rendering radius of outlier points far from the main sequence area is significantly reduced after outlier detection and rendering, and the perceptual quality of the point cloud sequence is improved.


(a) Andrew (b) David (c) Phil (d) Ricardo (e) Sarah (f) Head (g) House without roof (h) Egyptian mask (i) Facade (j) Frog

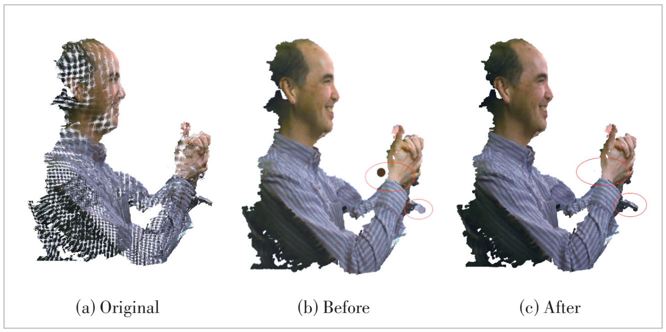
▲Figure 3. Point cloud sequence thumbnail

▼Table 1. Point cloud sequence details

Tested Point Clouds	Class (1: people 2: object)	Points	Geometry Precision/bit	Peak Value/bit
Andrew	1	1 276 312	10	1 023
David	1	1 492 780	10	1 023
Phil	1	1 089 091	10	1 023
Ricardo	1	2 592 758	10	1 023
Sarah	1	3 493 085	10	1 023
Head	2	13 903 516	12	4 095
House without roof	2	4 848 745	12	4 095
Egyptian mask	2	272 684	12	4 095
Facade	2	4 061 755	11	2 047
Frog	2	3 614 251	12	4 095


(a) Original (b) Before (c) After

▲Figure 4. Point cloud Head before and after optimization comparison diagram


(a) Original (b) Before (c) After

▲Figure 5. Point cloud Phil before and after optimization comparison diagram

We also compress the original point cloud sequence and calculate the PSNR using the original point rendering[10] and the proposed method. The improvement of PSNR using the proposed method is utilized to react to the effectiveness of rendering based on outlier detection.

In terms of point cloud compression, since outlier point detection does not target the attribute features of the point cloud sequence, the original color attributes are not compressed, and the original point cloud sequence attributes are kept lossless. Regarding geometric compression, we choose three geometric quantization parameters with large discretization, namely {1/10, 1/2, 15/18}, to verify the effectiveness of the proposed algorithm under different degrees of geometric compression in geometry point cloud compression (G-PCC).

To compute PSNR, we project each point cloud sequence to the three planes of the point cloud enclosing the box to form three views of the point cloud view. We render the original uncompressed sequence of the point cloud sequence through PccAppRendererV6.0 with the rendering radius set to 1.0 to obtain the reference view of the point cloud sequence. The quantized distortion images of each point cloud sequence before and after optimization are obtained in turn, and the mean squared error (MSE) and PSNR are calculated from the reference image and distortion image as shown in the following equations.

$$\mathrm{MSE} = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\left\| I(i,j) - K(i,j) \right\|^2, \tag{8}$$

$$\mathrm{PSNR} = 10\log_{10}\left(\frac{\mathrm{MAX}^2}{\mathrm{MSE}}\right), \tag{9}$$

where the image pixels are $m*n$, $I(i,j)$ represents the reference point cloud sequence view pixel points, $K(i,j)$ represents the quantized distortion point cloud sequence view pixel points, and MAX denotes the possible maximum pixel value of the image.

The improvement ratios of PSNR of the proposed method are shown in Table 2.

As can be seen from Table 2, the PSNR of each point cloud sequence has been improved using the proposed method regardless of the size of the geometric quantization parameters. As the geometric quantization parameter increases, more points are retained in the point cloud sequence. Therefore, for the same point cloud sequence, the advantage of the proposed method will be more significant with the increase in the density of the point cloud sequence. The reasons for the above phenomenon are that sparse point clouds are more likely to produce outliers, and the original point rendering will have worse subjective perception quality so that the PSNR improvement ratio will be more pronounced. After rendering through outlier detection, the PSNR of each point cloud sequence is improved. Specifically, the average PSNR improvement ratio is 3.6%.

3.3 Time Complexity

Specifically, we record the rendering time in the experiment in Section 3.2. Then we calculate the ratio of increased rendering time spent for the original rendering and the proposed rendering, as shown in Eq. (10).

$$T_r = \frac{T_1 - T_0}{T_0} \times 100\%, \tag{10}$$

▼**Table 2. PSNR improvement ratio before and after optimization with geometric quantization parameters**

Tested Point Clouds	Average Improvement Ratio of PSNR/%	Geometric Quantization Parameters/%		
		1/10	1/2	15/18
Andrew	+3.1	+1.9	+3.2	+4.1
David	+5.1	+2.6	+5.0	+7.6
Phil	+5.2	+2.7	+5.2	+7.7
Ricardo	+3.6	+3.2	+3.6	+4.1
Sarah	+4.4	+2.1	+4.9	+6.2
Facade	+1.5	+1.4	+1.4	+1.7
Head	+4.6	+3.7	+5.0	+5.3
House without roof	+6.0	+3.0	+7.2	+7.9
Egyptian mask	+0.8	+0.7	+0.8	+0.9
Frog	+2.0	+2.1	+2.1	+1.9

PSNR: peak signal-to-noise ratio

where T_r is the ratio of increased rendering time, T_1 is the rendering time using the proposed method, and T_0 is the original rendering time. The calculation results are shown in Table 3.

Observing the results, we can conclude that the rendering time is increased by 5.8% on average due to the added outlier detection.

3.4 Ablation Studies

To evaluate the effectiveness of using the local and global geometric features in outlier detection, we conduct studies using three-point cloud sequences: Andrew, David and Phil. For quantization, we still use the same geometric quantization parameters as set in Section 3.2: {1/10,1/2,15/18}.

The local and global geometric features are respectively used and then compared with the original rendering without optimization. For performance evaluation, the individual point cloud sequences are also projected onto the three planes of the point cloud enclosing the box to get the view from the corresponding perspective. The PSNR of the point cloud sequences under each rendering method is calculated from the undistorted point cloud sequence image and distorted point cloud sequence image. Finally, we calculate the average PSNR improvement compared with the original rendering method, as shown in Table 4.

It is seen that both local and global geometric features contribute to improving the quality of the point cloud sequence. However, both individual features are less effective than using both. In addition, the local geometric feature contributes more compared with the global geometric feature.

▼**Table 3. Ratio of increased rendering time**

Tested Point Clouds	Ratio of Increased Rendering Time (T_r)/%
Andrew	6
David	5
Phil	6
Ricardo	5
Sarah	5
Facade	6
Head	8
House without roof	6
Egyptian mask	5
Frog	6
Average	**5.8**

▼**Table 4. Average improvement ratio of PSNR in ablation studies**

Tested Point Clouds	PSNR Average Improvement Ratio/%		
	Global only	Local only	Global + local
Andrew	+0.6	+1.2	+3.1
David	+0.9	+2.1	+5.1
Phil	+1.0	+2.4	+5.2

PSNR: peak signal-to-noise ratio

4 Conclusions

In this paper, we focus on optimizing the perceptual quality of point cloud rendering through outlier detection. We evaluate the performance of the proposed method in terms of perceptual quality and the PSNR improvement ratio. In addition, we evaluate time complexity analysis and perform ablation studies. Future work may include applying other methods to outlier detection in rendering to improve the perceptual quality after point cloud rendering.

References

[1] SEUFERT M, KARGL J, SCHAUER J, et al. Different points of view: impact of 3D point cloud reduction on QoE of rendered images [C]//The 12th International Conference on Quality of Multimedia Experience. IEEE, 2020: 1 – 6. DOI: 10.1109/QoMEX48832.2020.9123143

[2] DUMIC E, BATTISTI F, CARLI M, et al. Point cloud visualization methods: a study on subjective preferences [C]//The 28th European Signal Processing Conference. IEEE, 2020: 595 – 599

[3] ZHAO X, ZHANG B W, WU J J, et al. Relationship-based point cloud completion [J]. IEEE transactions on visualization and computer graphics, 2022, 28 (12): 4940 – 4950. DOI: 10.1109/TVCG.2021.3109392

[4] CHEN H H, WEI M Q, SUN Y X, et al. Multi-patch collaborative point cloud denoising via low-rank recovery with graph constraint [J]. IEEE transactions on visualization and computer graphics, 2020, 26(11): 3255 – 3270. DOI: 10.1109/TVCG.2019.2920817

[5] LI H Q, LI L, LI Z. A review of point cloud compression [J]. ZTE technology journal, 2021, 27(1): 5 – 9. DOI: 10.12142/ZTETJ.202101003

[6] LIU Q, SU H L, DUANMU Z F, et al. Perceptual quality assessment of colored 3D point clouds [J]. IEEE transactions on visualization and computer graphics, 2023, 29(8): 3642 – 3655. DOI: 10.1109/TVCG.2022.3167151

[7] LIU Q, YUAN H, HOU J H, et al. Model-based joint bit allocation between geometry and color for video-based 3D point cloud compression [J]. IEEE transactions on multimedia, 2021, 23: 3278 – 3291. DOI: 10.1109/TMM.2020.3023294

[8] VAN DER HOOFT J, VEGA M T, TIMMERER C, et al. Objective and subjective QoE evaluation for adaptive point cloud streaming [C]//The 12th International Conference on Quality of Multimedia Experience. IEEE, 2020: 1 – 6. DOI: 10.1109/QoMEX48832.2020.9123081

[9] PHARR M, JAKOB W, HUMPHREYS G. Physically based rendering: from theory to implementation [M]. Massachusetts, USA: MIT Press, 2023

[10] JAVAHERI A, BRITES C, PEREIRA F, et al. Point cloud rendering after coding: impacts on subjective and objective quality [J]. IEEE transactions on multimedia, 2021, 23: 4049 – 4064. DOI: 10.1109/TMM.2020.3037481

[11] XU I L, YANG Q, YANG D, et al. Challenges and key technologies of point cloud quality evaluation [J]. Journal of Communication University of China: natural science edition, 2021, 28(5): 11

[12] KAZHDAN M, BOLITHO M, HOPPE H. Poisson surface reconstruction [C]//The Fourth Eurographics Symposium on Geometry Processing. ACM, 2006

[13] FENG Y P, ZHONG H X, PANG Y J. Non-mesh rendering based on points [C]//International Conference on Machine Learning and Cybernetics. IEEE, 2005: 5442 – 5446. DOI: 10.1109/ICMLC.2005.1527906

[14] XUE A R. Research on spatial outlier excavation technology [D]. Zhenjiang: Jiangsu Univeristy, 2009

[15] MANDHARE H C, IDATE S R. A comparative study of cluster based outlier detection, distance based outlier detection and density based outlier detection techniques [C]//International Conference on Intelligent Computing and Control Systems (ICICCS). IEEE, 2018: 931 – 935. DOI: 10.1109/ICCONS.2017.8250601

[16] LIN R H, HU H, WEN Z K, et al. Research on denoising and segmentation algorithm application of pigs' point cloud based on DBSCAN and PointNet [C]//IEEE International Workshop on Metrology for Agriculture and Forestry (MetroAgriFor). IEEE, 2021: 42 – 47. DOI: 10.1109/MetroAgriFor52389.2021.9628501

[17] XU S Y, LIU H Y, DUAN L T, et al. An improved LOF outlier detection algorithm [C]//IEEE International Conference on Artificial Intelligence and Computer Applications. IEEE, 2021: 113 – 117. DOI: 10.1109/ICAICA52286.2021.9498181

[18] MOSALLAM B E, AHMED S H. Exploring effective outlier detection in IoT: a systematic survey of techniques and applications [C]//Intelligent Methods, Systems, and Applications (IMSA). IEEE, 2023: 375 – 380. DOI: 10.1109/IMSA58542.2023.10255071

[19] XUE A R, YAO L, JU, T, et al. A review of outlier mining methods [J]. Computer science, 2008, 35(11): 13 – 18

[20] ZHAO P. Outlier detection and model reconstruction of 3D point cloud data [D]. Dalian: Dalian University of Technology, 2015

[21] LIU R, WAN W G, ZHOU Y Y, et al. Normal estimation algorithm for point cloud using KD-Tree [C]//IET International Conference on Smart and Sustainable City. IET, 2013. DOI: 10.1049/cp.2013.1978

Biographies

YIN Yujie (yinyujie@mail.nwpu.edu.cn) received his BS degree in electronic information engineering from Hohai University, China in 2023, and he is currently pursuing his MS degree in information and communication engineering from Northwestern Polytechnical University, China. His main research interests are point clouds and video coding.

CHEN Zhang received his BS degree in electrical and information engineering and MS degree in signal and information processing from Northwestern Polytechnical University, China. Currently, he is working on his PhD degree at Northwestern Polytechnical University, and his main research interests are point cloud compression and point cloud quality assessment.

Local Scenario Perception and Web AR Navigation

SHI Wenzhe[1,2], LIU Yanbin[1,2], ZHOU Qinfen[1]

(1. State Key Laboratory of Mobile Network and Mobile Multimedia Technology, Shenzhen 518055, China;
2. ZTE Corporation, Shenzhen 518057, China)

DOI: 10.12142/ZTECOM.202304007

https ://kns.cnki.net/kcms/detail/34.1294.TN.20231129.1522.004.html, published online November 30, 2023

Manuscript received: 2023-06-20

Abstract: This paper proposes a local point cloud map-based Web augmented reality (AR) indoor navigation system solution. By delivering the local point cloud map to the web front end for positioning, the real-time positioning can be implemented only with the help of the computing power of the web front end. In addition, with the characteristics of short time consumption and accurate positioning, an optimization solution to the local point cloud map is proposed, which includes specific measures such as descriptor de-duplicating and outlier removal, thus improving the quality of the point cloud. In this document, interpolation and smoothing effects are introduced for local map positioning, enhancing the anchoring effect and improving the smoothness and appearance of user experience. In small-scale indoor scenarios, the positioning frequency on an iPhone 13 can reach 30 fps, and the positioning precision is within 50 cm. Compared with an existing mainstream visual-based positioning manner for AR navigation, this specification does not rely on any additional sensor or cloud computing device, thereby greatly saving computing resources. It takes a very short time to meet the real-time requirements and provide users with a smooth positioning effect.

Keywords: Web AR; three-dimensional reconstruction; navigation; positioning

Citation (Format 1): SHI W Z, LIU Y B, ZHOU Q F. Local scenario perception and Web AR navigation [J]. ZTE Communications, 2023, 21(4): 54 – 59. DOI: 10.12142/ZTECOM.202304007

Citation (Format 2): W. Z. Shi, Y. B. Liu, and Q. F. Zhou, "Local scenario perception and Web AR navigation," *ZTE Communications*, vol. 21, no. 4, pp. 54 – 59, Dec. 2023. doi: 10.12142/ZTECOM.202304007.

1 Introduction

There are three existing positioning manners for augmented reality (AR) navigation: visual positioning, GPS positioning, and multi-sensor fusion positioning. Specifically, visual positioning is based on a conventional simultaneous localization and mapping (SLAM) framework, and the main steps include feature extraction, feature matching, and position solving. This model has high precision, but requires high computing power and cannot run on the Web side.

Outdoor AR navigation can achieve good results by using GPS technology. However, in an indoor scenario, the signal strength of GPS is greatly affected, and consequently, positioning precision is obviously reduced. Therefore, the GPS-based positioning manner cannot be applied to indoor AR navigation.

A positioning manner of multi-sensor fusion is to obtain the position of a camera by fusing data from sensors such as an inertial sensor, a laser radar, Bluetooth, and Wi-Fi. In this manner, although positioning accuracy is high, the sensor is vulnerable to environments, thereby decreasing positioning performance. In addition, in this manner, a large number of sensors need to be calibrated and fused, and a development cost is high.

Although the foregoing three methods can obtain relatively high precision in certain specific scenarios, they cannot be applied to an indoor Web AR navigation scenario. The core reason is that the computing power on the Web side is limited and cannot meet the intensive requirements of AR computing. If a positioning system that meets performance requirements can be implemented on the web front end by using limited computing power, dependence on an external computing environment or device will be reduced, development costs can be cut, and the application scope and user experience of indoor Web AR navigation will be greatly improved.

2 Key Technologies of Visual Perception

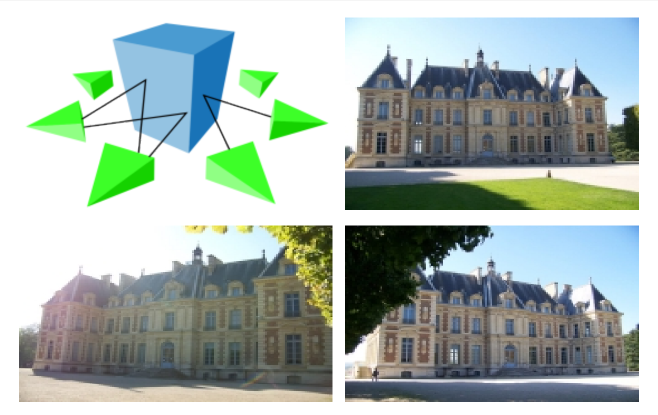
2.1 3D Reconstruction Techniques

To implement a visual method of good positioning, precision is indispensable for a robust three-dimensional reconstruction process. An objective of three-dimensional reconstruction is to obtain a geometric structure and a structure of an object or a scene from a group of images, which may be implemented by using a motion recovery structure (Structure-from-Motion, SFM). SFM is a method for implementing three-dimensional reconstruction and mainly used in a phase of con-

This work is supported in part by ZTE Industry-University-Institute Cooperation Funds.

structing a sparse point cloud image in the three-dimensional reconstruction. A complete three-dimensional reconstruction process generally uses a Multi-View Stereo (MVS) algorithm to implement dense reconstruction. As shown in Fig. 1, SFM is mainly used for creating diagrams and restoring the structure of the scenario. According to the difference of image data processing flows, SFM can be divided into four types: incremental SFM, global SFM, distributed SFM, and hybrid SFM. The latter two types are usually used to resolve a very large-scale data scenario and are based on the former two types. Incremental SFM can be divided into two steps. The first step is to find the initial correspondence to extract robust and well-distributed features to match the image pairs, and the second step is to implement incremental reconstruction to estimate image position and 3D structure by image registration, triangulation, bundle adjustment (BA), and abnormal value removal. The initial corresponding abnormal value needs to be removed through geometric verification. Generally, when the number of restored image frames accounts for a certain proportion, global BA is performed. Because of the incremental processing of BAs, the precision of the incremental SFM is usually relatively high and the robustness is relatively good. However, with the increase of the images, the processing scale of the BAs becomes larger and larger. Therefore, there are also disadvantages such as low efficiency and large memory usage. In addition, the incremental SFM also has the problem of accumulative drift because of the incremental addition of images. Typical SFM frameworks include Bundler and COLMAP.

CAO et al. [1] proposed a fast and robust feature-tracking method for 3D reconstruction using SFM. First, to reduce calculation costs, a large number of image sets are clustered into some small image sets by using a feature clustering method to avoid some incorrect feature matching. Second, a joint search set method is used to implement fast feature matching, which may further save calculation time of feature tracking. Third, a geometric constraint method is proposed to remove an abnormal value from a track generated by a feature tracking method. This method can deal with the influence of image distortion, scale change and illumination change. LINDENBERGER et al.[2] directly align low-level image information from multiple views, optimize feature point positions using depth feature metrics after feature matching, and perform BA during incremental reconstruction using similar depth feature metrics. In this process, an image-dense feature map is first extracted by using a convolution network, two-dimensional observation of the same three-dimensional point in different images is obtained using sparse feature matching, the location of a corresponding feature point in the image is adjusted, SFM reconstruction is performed according to the adjusted location, and a residual of SFM optimization in the reconstruction process changes from a reprojection error to a depth feature measurement error. This improvement is robust to large-scale detection of noise and appearance changes because it optimizes feature measurement errors for dense features based on neural network prediction.



▲Figure 1. Shooting a panoramic video of the scene

Some accumulated drift problems are solved through global SFM. In an image matching process, a basic/essential matrix between images is obtained, and relative rotation and relative translation between the images may be obtained by means of decomposition. Global rotation can be restored by using relative rotation as a constraint. Global panning can then be restored using the global rotation and relative panning constraints. Because the number of times of building and optimizing global BA is small, the efficiency of global SFM is high. However, it is difficult to solve the translation average because the relative translation constraint only constrains the translation direction and the scale is unknown. In addition, the translation average solving process is sensitive to external points. Therefore, in actual applications, the global SFM is limited.

2.2 Space Visual Matching Technology

How to extract robust, accurate and sufficient image correspondence is the key problem of 3D reconstruction. With the development of deep learning, the image matching methods based on learning achieve excellent performance. A typical image matching process is divided into three steps: feature extraction, feature description, and feature matching.

Detection methods based on deep convolution networks search for points of interest by building response maps, including the supervisory method[3–4], self-supervised method[5–6], and unsupervised method[7–8]. The supervisory approach uses an anchor to guide the training process of a model, but the performance of the model is likely limited by the anchor construction approach. Self-supervised and unsupervised methods do not require manual annotation of data, and they focus on geometric constraints between image pairs.

The feature descriptor uses local information around the point of interest to establish a correct correspondence between image features. Due to the ability of information extraction and representation, depth techniques have also performed well in the description of features. The feature description problem based on deep learning is usually a supervised learning problem, that is, learning a representation that makes matched features in the measurement space as close as possible and unmatched features as far as possible[9]. Learning-based descriptors largely avoid the need for human experience and prior knowledge. An existing feature description method based on learning is classified into two types: measurement learning[10 – 11] and descriptor learning[12 – 13]. A difference lies in the output content of a descriptor.

Metric learning methodology is used for metric discrimination for similarity measurement, while descriptor learning generates descriptor representations from the original image or image block. In these methods, SuperGlue[14] is a network that can perform feature matching and filter out extrinsic points at the same time, where feature matching is implemented by solving a differential optimization transfer problem, a loss function is constructed by using a graph neural network, and a flexible content aggregation mechanism is proposed based on an attention mechanism. Therefore, SuperGlue can simultaneously sense a potential three-dimensional scene and perform feature matching. LoFTR[15] uses transformer modules with a self-attentive layer and a cross-attentive layer to process dense local features extracted from the convolutional network by first extracting dense matches at low feature resolution (1/8 of the image dimension) and then selecting the matches with high confidence from those matches using a relevant method to refine them to a high-resolution sub-pixel level. In this way, the large acceptance field of the model enables the converted signature to reflect the context and location information, and the matching is implemented through multiple layers of self-attention and cross-attention. Many methods integrate feature detection, feature description, and feature matching into the matching pipeline in an end-to-end manner, which helps improve matching performance.

Visual orientation is a problem of estimating a 6-DoF camera pose from which a given image is taken relative to a reference scene representation. The classical approach to visual positioning is structure-based, meaning that they rely on the 3D reconstruction of the environment (that is, point clouds) and use local feature matching to establish a mapping relationship between the query image and 3D map. Image retrieval can be used to reduce the search space by only considering the most similar reference images rather than all possibilities. Another approach is to interpolate or estimate the relative posture between the queried and retrieved reference images directly from the reference images, which is independent of the 3D reconstruction results. In the scene point regression method, a correspondence between a two-dimensional pixel position and a three-dimensional point may be directly determined by using a deep neural network (DNN), and the position of a camera is calculated similarly to a structure-based method. Modern scene regression benefits from 3D reconstruction during training but does not depend on it. Finally, the absolute posture regression method uses DNN end-to-end posture estimation. These approaches differ in generalization capabilities and location accuracy.

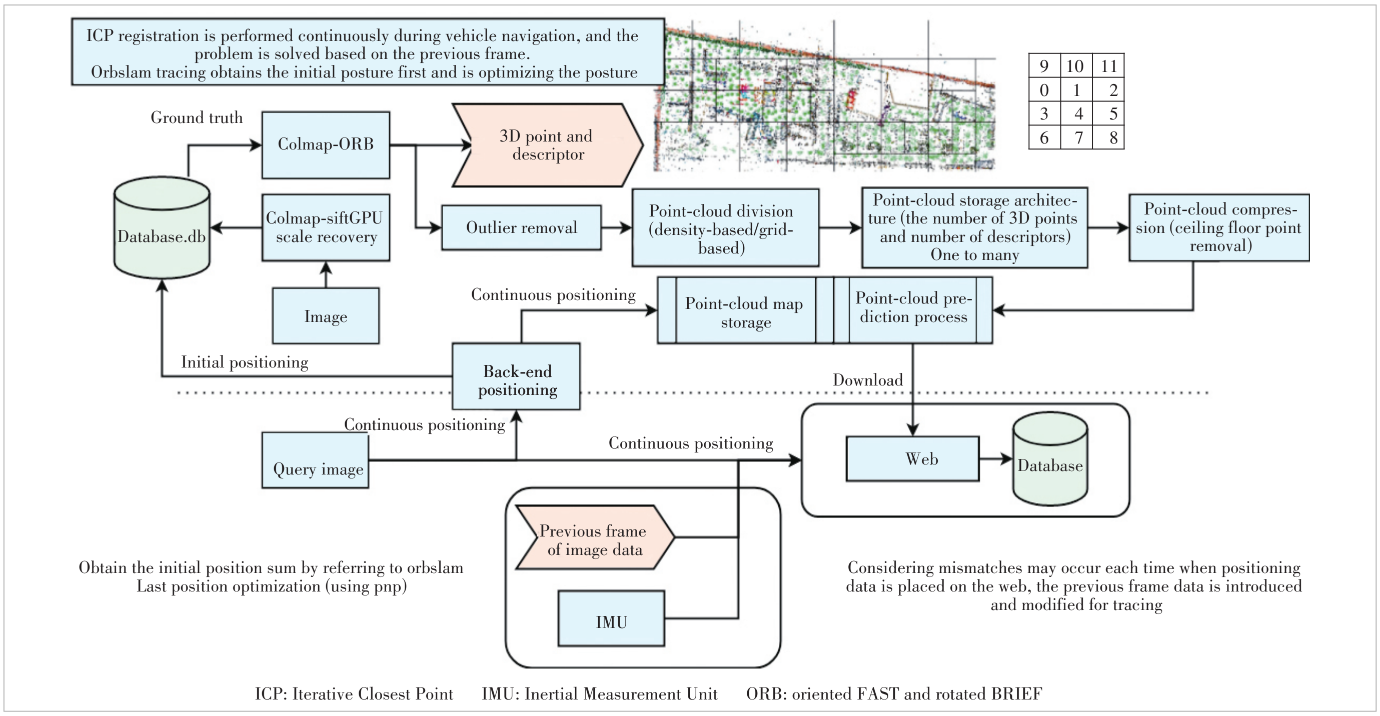
In addition, some methods rely on 3D reconstruction, while others only require reference images with position marks. The advantage of using a 3D reconstruction is that the position generated is very accurate, and the disadvantage is that these 3D reconstructions are sometimes difficult to obtain or even more difficult to maintain. For example, if the environment changes, the position needs to be updated. For classical structure-based work, reference may be made to a general visual positioning framework proposed by SARLIN et al.[16] The framework can predict both local features and global descriptors by using a hierarchical positioning method, so as to implement accurate 6-DoF positioning. Using a coarse-to-fine localization pattern, the method first performs a global search to obtain location assumptions and then matches local features in these candidate locations. This layered approach saves uptime for real-time operations. This method presents a hierarchical feature network (HF-Net), which jointly estimates local and global features, shares computation to the maximum extent, and uses a multi-task still compression model.

3 Web AR Navigation System Based on Local Scenario Perception

This paper presents an indoor Web AR navigation system architecture based on the local point cloud map. By delivering the space local point cloud map to the web front end for positioning, the real-time positioning can be implemented only by using the computing power of the web front end, which has the characteristics of short time consumption and accurate positioning. In addition, this paper proposes an optimization solution to the local point cloud map, including specific measures such as descriptor deduplication and outlier elimination, which improves the quality of the point cloud. Finally, interpolation and smoothing effects are introduced to local map localization to enhance an anchoring effect and improve smoothness and appearance of user experience. In a small-scale indoor scenario, a localization frequency on an iPhone 13 may reach 30 fps, and localization precision is within 50 cm. In this paper, a function of implementing real-time positioning by using only Web front-end computing power is proposed for the first time. It outperforms existing mainstream visual-based positioning for AR navigation, GPS-based positioning, and multi-sensor fusion positioning. The proposed method can significantly save computing resources without the help of any additional sensors or cloud computing devices. It takes a very short time to meet the real-time requirements and provide users with smooth positioning, improving user experience.

Fig. 2 shows the proposed Web AR indoor navigation system based on local point cloud map positioning. This system consists of three modules: offline map creation, server, and web.

The offline map creation module is mainly responsible for the reconstruction of a point cloud map. Three-dimensional reconstruction is implemented by photographing an environmental image that needs to be reconstructed and then scale-based


ICP registration is performed continuously during vehicle navigation, and the problem is solved based on the previous frame.
Orbslam tracing obtains the initial posture first and is optimizing the posture
Ground truth
Colmap-ORB
3D point and descriptor
Database.db
Colmap-siftGPU scale recovery
Outlier removal
Point-cloud division (density-based/grid-based)
Point-cloud storage architecture (the number of 3D points and number of descriptors) One to many
Point-cloud compression (ceiling floor point removal)
Image
Continuous positioning
Point-cloud map storage
Point-cloud prediction process
Initial positioning
Back-end positioning
Download
Continuous positioning
Query image
Continuous positioning
Web
Database
Previous frame of image data
IMU
Obtain the initial position sum by referring to orbslam
Last position optimization (using pnp)
Considering mismatches may occur each time when positioning data is placed on the web, the previous frame data is introduced and modified for tracing

9	10	11
0	1	2
3	4	5
6	7	8

ICP: Iterative Closest Point IMU: Inertial Measurement Unit ORB: oriented FAST and rotated BRIEF

▲Figure 2. Local scenario perception and the proposed web navigation system architecture

restoration is performed, to finally obtain a sparse point cloud map and save the sparse point cloud map in the format of a 3D point plus a descriptor. Then, the point cloud is visualized and divided according to the preset interest point when the user wants to perform a model anchoring display. The related geo-fence information is set, which is mainly used for service experience after entering the local point cloud range. Currently, the geo-fence range is mainly 3 m – 5 m and established according to a specific scenario. The sparse point cloud is divided into multiple local point cloud maps. Next, the point cloud is optimized by using descriptor deduplication and outlier removal and is stored in the bin format.

The server performs positioning on the captured initial positioning picture to obtain an initial positioning posture of the camera, so as to determine a local point cloud closest to an initial positioning point position. The local point cloud communication service is responsible for delivering the specific local point cloud to the Web front end in accordance with the request of the Web front end for the local point cloud.

The Web front end sends a request to the server end by using a local point cloud communication service, receives specific local point cloud information, and then captures an image of a video frame by using a local point cloud positioning system of the Web front end. The Web front end obtains corresponding camera position information for positioning and then renders a navigation route and a corresponding material based on the camera position information obtained by positioning. Fig. 2 shows how to implement AR navigation through the local cloud.

The time consumption of each step of the local point cloud positioning algorithm is collected and optimized, including image data transmission on a Web end, improvement of a feature extraction algorithm, feature matching optimization, etc. Table 1 shows the performance test of cloud positioning of different models at different point sizes. In the point cloud of 0.9 MB, the optimized algorithm can reach 91 fps on an iPhone 13.

The redundancy of the point cloud size greatly affects the accuracy of the local point cloud positioning algorithm. Therefore, two local point cloud optimization solutions are designed: 1) Using filter feature descriptors to remove duplicates (Fig. 3); 2) Using test data to filter real and valid point cloud data and remove redundancy.

Considering that the positioning algorithm is an optimization problem (reducing a reprojection error), it is extremely affected by noise, and therefore a final obtained track is not smooth

▼Table 1. Different mobile phone models in frames per second

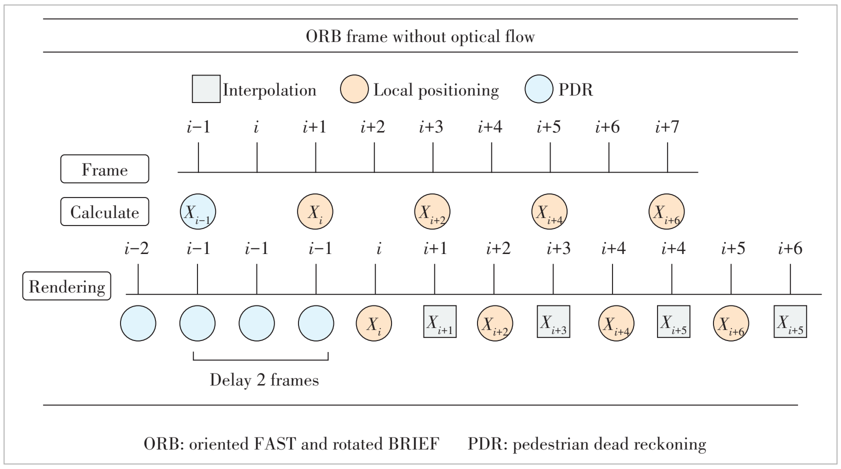
Descriptor Size/MB	Mobile Phone Model	Extracting ORB Features	Feature Matching (KNN)	PNP	Total Calculated Time/ms	Frames per Second/fps
0.9	MEIZU 11	1.052	38	4	50	20
	OnePlus 6	0.876	37	4	46	21
	Xiaomi 11	0.557	19	2	26	38
	Iphone 13	0.098	8	2	11	90

KNN: k-Nearest Neighbor ORB: oriented FAST and rotated BRIEF PNP: Perspective-n-Points enough. To ensure a stable anchoring effect, a filtering manner is used to optimize the positioning algorithm as follows.

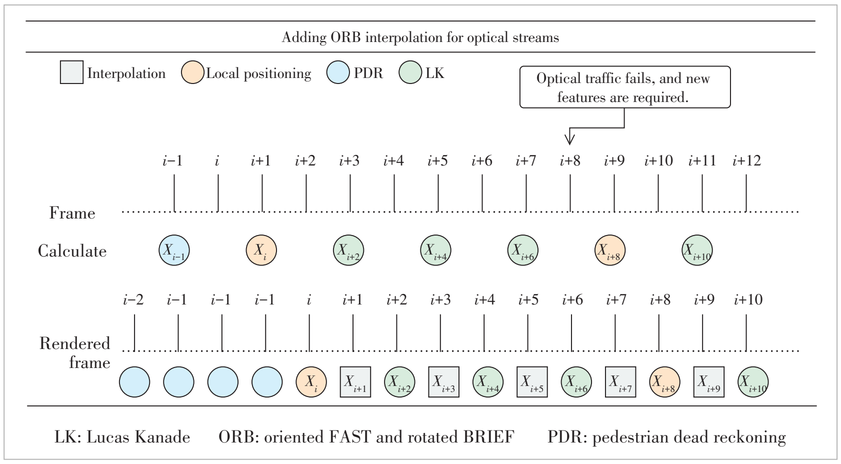
1) A high-pass filter and a low-pass filter are used to eliminate incorrect positioning (Fig. 4);

2) The camera position information of the first k frames and the sliding average value are used to smooth the track of the current frame.

Because the original environment on the web side supports only a single thread and the location algorithm based on the local point cloud cannot meet the real-time requirement (20 fps) with limited computing resources on the web side, the proposed algorithm is optimized by delaying video frames (Fig. 5). To ensure the stability of the local point cloud positioning algorithm, the optical flow tracking algorithm is introduced to improve the number of 2D-3D matches by using the previous frame's prior knowledge, as shown in Fig. 6. Figs 7 and 8 show the experimental results without and with the optical flow respectively. The stability of model anchoring is improved during the positioning process.


ORB frame without optical flow
Interpolation
Local positioning
PDR
Frame
Calculate
Rendering
i−1 i i+1 i+2 i+3 i+4 i+5 i+6 i+7
X_{i-1} X_i X_{i+2} X_{i+4} X_{i+6}
i−2 i−1 i−1 i−1 i i+1 i+2 i+3 i+4 i+4 i+5 i+6
X_i X_{i+1} X_{i+2} X_{i+3} X_{i+4} X_{i+5} X_{i+6} X_{i+5}
Delay 2 frames
ORB: oriented FAST and rotated BRIEF PDR: pedestrian dead reckoning

▲**Figure 5. Delaying two frames without optical flow**

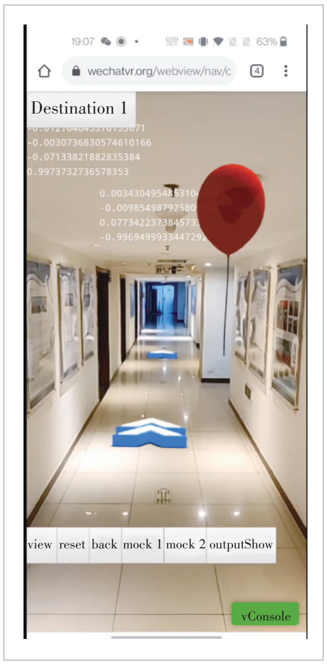

Adding ORB interpolation for optical streams
Interpolation
Local positioning
PDR
LK
Optical traffic fails, and new features are required.
i−1 i i+1 i+2 i+3 i+4 i+5 i+6 i+7 i+8 i+9 i+10 i+11 i+12
Frame
Calculate
X_{i-1} X_i X_{i+2} X_{i+4} X_{i+6} X_{i+8} X_{i+10}
i−2 i−1 i−1 i−1 i i+1 i+2 i+3 i+4 i+5 i+6 i+7 i+8 i+9 i+10
Rendered frame
X_i X_{i+1} X_{i+2} X_{i+3} X_{i+4} X_{i+5} X_{i+6} X_{i+7} X_{i+8} X_{i+9} X_{i+10}
LK: Lucas Kanade ORB: oriented FAST and rotated BRIEF PDR: pedestrian dead reckoning

▲**Figure 6. Experiment of adding interpolation for optical streams**

4 Conclusions

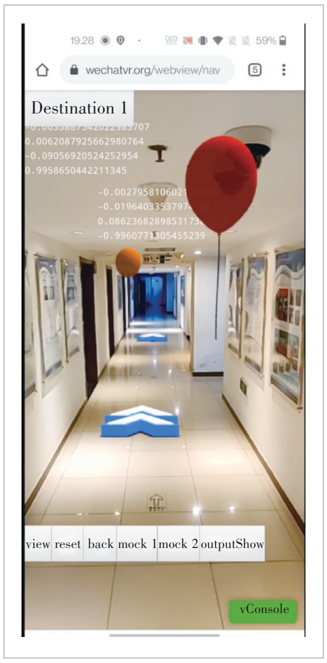
This paper proposes a Web AR indoor navigation system based on local point cloud map positioning, which has beneficial effects on technical value compared with the prior art. It innovatively proposes the idea of point-cloud distribution, that is, to download the map of local point-cloud to the Web front end and use the computing power of the Web front end for positioning. Compared with an existing mainstream visual-based positioning manner for AR navigation, GPS-based positioning manner and multi-sensor fusion positioning manner, the positioning manner provided in the present invention does not depend on any additional sensor or external computing environment, thereby reducing development and deployment costs.

A lightweight web front-end positioning algorithm is presented for indoor Web AR navigation when the computational power of the web front-end is limited. A degree of dependence on network communication is reduced, the requirement of Web AR navigation on a network environment is reduced, and environment adaptability is improved. It takes a short time to deliver the web front-end positioning system to the point cloud for indoor Web AR navigation. In a small-scale indoor scenario, the positioning frequency on the iPhone 13 can reach 90 fps, which brings users a smooth user experience based on satisfying the real-time positioning requirements for Web AR navigation.



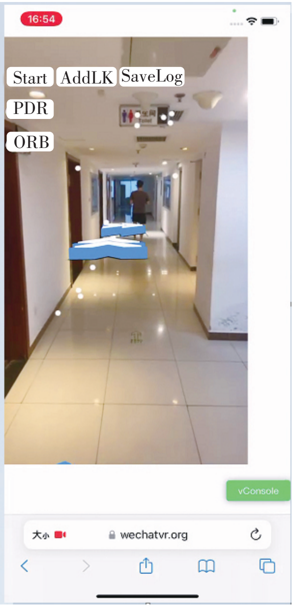
```
-0.0121640455376133671
-0.0030736830574610166
-0.07133821882835384
0.9973732736578353
0.00343049548531040
-0.00985498792580
0.077342237384573
-0.996949993344729
```

▲**Figure 3. No filter is added for Web AR navigation**



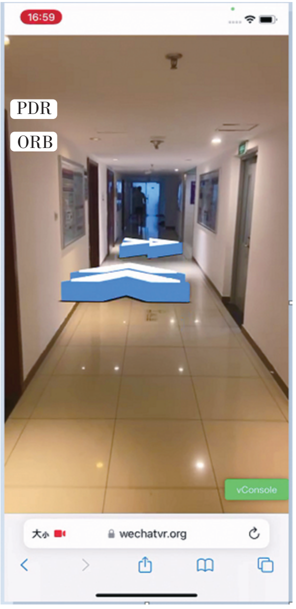
```
-0.0055887542022985707
0.0062087925662980764
-0.09056920524252954
0.9958650442211345
-0.0027958106021
-0.019640335379?
0.0862368289853173
-0.9960771305455239
```

▲**Figure 4. High-pass filter is added for Web AR navigation**



ORB: oriented FAST and rotated BRIEF
PDR: pedestrian dead reckoning

▲ **Figure 7. The proposed algorithm is optimized without optical flow**



ORB: oriented FAST and rotated BRIEF
PDR: pedestrian dead reckoning

▲ **Figure 8. Experiment diagram of Interpolation + Optical Flow**

Acknowledgement

We would like to acknowledge HUANG Yakun and LIU Jiarui for their contribution to this paper.

References

[1] CAO M W, JIA W, LV Z H, et al. Fast and robust feature tracking for 3D reconstruction [J]. Optics & laser technology, 2019, 110: 120 – 128. DOI: 10.1016/j.optlastec.2018.05.036

[2] LINDENBERGER P, SARLIN P E, LARSSON V, et al. Pixel-perfect structure-from-motion with featuremetric refinement [C]//Proc. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2022: 5967 – 5977. DOI: 10.1109/ICCV48922.2021.00593

[3] YI K M, TRULLS E, LEPETIT V, et al. LIFT: learned Invariant Feature Transform [C]//European Conference on Computer Vision. Springer, 2016: 467 – 483. DOI: 10.1007/978-3-319-46466-4_28

[4] ZHANG X, YU F X, KARAMAN S, et al. Learning discriminative and transformation covariant local feature detectors [C]//Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017: 4923 – 4931. DOI: 10.1109/CVPR.2017.523

[5] ZHANG L G, RUSINKIEWICZ S. Learning to detect features in texture images [C]//Proc. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2018: 6325 – 6333. DOI: 10.1109/CVPR.2018.00662

[6] DETONE D, MALISIEWICZ T, RABINOVICH A. SuperPoint: self-supervised interest point detection and description [C]//Proc. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, 2018: 337 – 33712. DOI: 10.1109/CVPRW.2018.00060

[7] LAGUNA A B, RIBA E, PONSA D, et al. Key.Net: keypoint detection by handcrafted and learned CNN filters [C]//Proc. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2020: 5835 – 5843. DOI: 10.1109/ICCV.2019.00593

[8] ONO Y, TRULLS E, FUA P, et al. LF-net: learning local features from images [C]//Proc. 32nd International Conference on Neural Information Processing Systems. NIPS, 2018: 6273 – 6247. DOI:10.5555/3327345.3327521

[9] SCHÖNBERGER J L, HARDMEIER H, SATTLER T, et al. Comparative evaluation of hand-crafted and learned local features [C]//Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017: 6959 – 6968. DOI: 10.1109/CVPR.2017.736

[10] WANG J, ZHOU F, WEN S L, et al. Deep metric learning with angular loss [C]//Proc. 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, 2017: 2612 – 2620. DOI: 10.1109/ICCV.2017.283

[11] ZAGORUYKO S, KOMODAKIS N. Learning to compare image patches via convolutional neural networks [C]//Proc. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2015: 4353 – 4361. DOI: 10.1109/CVPR.2015.7299064

[12] LUO Z X, SHEN T W, ZHOU L, et al. ContextDesc: local descriptor augmentation with cross-modality context [C]//Proc. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020: 2522 – 2531. DOI: 10.1109/CVPR.2019.00263

[13] TIAN Y R, YU X, FAN B, et al. SOSNet: second order similarity regularization for local descriptor learning [C]//Proc. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020: 11008 – 11017. DOI: 10.1109/CVPR.2019.01127

[14] SARLIN P E, DETONE D, MALISIEWICZ T, et al. SuperGlue: learning feature matching with graph neural networks [C]//Proc. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020: 4937 – 4946. DOI: 10.1109/CVPR42600.2020.00499

[15] SUN J M, SHEN Z H, WANG Y A, et al. LoFTR: detector-free local feature matching with transformers [C]//Proc. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2021: 8918 – 8927. DOI: 10.1109/CVPR46437.2021.00881

[16] SARLIN P E, CADENA C, SIEGWART R, et al. From coarse to fine: Robust hierarchical localization at large scale [C]//Proc. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020: 12708 – 12717. DOI: 10.1109/CVPR.2019.01300

Biographies

SHI Wenzhe (shi.wenzhe@zte.com.cn) is a strategy planner and engineer for XRExplore Platform product planning at ZTE Corporation. He is also a member of the National Key Laboratory for Mobile Network and Mobile Multimedia Technology. His research interests include indoor visual AR navigation, SFM 3D reconstruction, visual SLAM, real-time cloud rendering, VR, and spatial perception.

LIU Yanbin is a strategy planner and product manager for XRExplore Platform product planning at ZTE Corporation. He is also a member of the National Key Laboratory for Mobile Network and Mobile Multimedia Technology. His research interests include real-time remote rendering, visual SLAM, and computer vision.

ZHOU Qinfen is the XR product leader director of new media industry and a senior architect of ZTE Corporation. She has more than 20 years of experience in the communication industry and media business. She has held the positions of product manager of short message center, product manager of cloud desktop, product line cost director, and XR product director at ZTE Corporation. She has a thorough understanding of products and related standards including Short Message Center, Cloud Desktop GPU Virtualization, XR, etc. As a member of the Shenzhen 8K UHD Video Industry Cooperation Alliance (SUCA) and Virtual Display Professional Committee of Jiangsu Communication Association, she leads a team responsible for the research on the latest video technology and related standards.

Research on Fall Detection System Based on Commercial Wi-Fi Devices

GONG Panyin[1], ZHANG Guidong[1], ZHANG Zhigang[2,3], CHEN Xiao[2,3], DING Xuan[1]

(1. School of software, Tsinghua University, Beijing 100084, China;
2. ZTE Corporation, Shenzhen 518057, China;
3. State Key Laboratory of Mobile Network and Mobile Multimedia Technology, Shenzhen 518055, China)

DOI: 10.12142/ZTECOM.202304008

https://kns.cnki.net/kcms/detail/34.1294.TN.20231025.1621.002.html, published online October 26, 2023

Manuscript received: 2023-04-07

Abstract: Falls are a major cause of disability and even death in the elderly, and fall detection can effectively reduce the damage. Compared with cameras and wearable sensors, Wi-Fi devices can protect user privacy and are inexpensive and easy to deploy. Wi-Fi devices sense user activity by analyzing the channel state information (CSI) of the received signal, which makes fall detection possible. We propose a fall detection system based on commercial Wi-Fi devices which achieves good performance. In the feature extraction stage, we select the discrete wavelet transform (DWT) spectrum as the feature for activity classification, which can balance the temporal and spatial resolution. In the feature classification stage, we design a deep learning model based on convolutional neural networks, which has better performance compared with other traditional machine learning models. Experimental results show our work achieves a false alarm rate of 4.8% and a missed alarm rate of 1.9%.

Keywords: fall detection; commercial Wi-Fi devices; discrete wavelet transform; deep learning model

Citation (Format 1): GONG P Y, ZHANG G D, ZHANG Z G, et al. Research on fall detection system based on commercial Wi-Fi devices [J]. ZTE Communications, 2023, 21(4): 60 – 68. DOI: 10.12142/ZTECOM.202304008

Citation (Format 2): P. Y. Gong, G. D. Zhang, Z. G. Zhang, et al., “Research on fall detection system based on commercial Wi-Fi devices,” *ZTE Communications*, vol. 21, no. 4, pp. 60 – 68, Dec. 2023. doi: 10.12142/ZTECOM.202304008.

1 Introduction

Population aging is a common phenomenon in the world today, and the health as well as safety of the elderly is a growing concern. Every year, one-third of the elderly people over 65 fall down, resulting in injury or even death[1]. Most of the elderly deaths caused by falls are due to untimely treatment, so fall detection has become an important method to prevent fall-related deaths in the elderly.

Extensive research has been conducted on fall detection techniques. Traditional methods mainly use cameras[2 – 3], wearable sensors[4 – 5], and ambient environment sensing-based approaches[6 – 8] for fall detection. Camera-based detection systems require infrastructure deployment and video data collection, which raises the user’s privacy concerns. Solutions based on wearable sensors require users to carry specific devices, which is inconvenient. Ambient environment sensing-based approaches require specific infrastructures (e.g., radar and infrared) and are expensive, which prevents them from pervasive applications. Therefore, it is important to find a fall detection solution that is device-independent, privacy protecting, secure, inexpensive and easy to deploy. The emergence of commercial Wi-Fi[9 – 10], which can effectively meet these conditions, has also received a lot of attention from researchers. Wi-Fi devices sense the user’s activity by analyzing the channel state information (CSI) of the received signal, which enables various applications such as gesture recognition[11 – 12], gait recognition[13 – 14], and trajectory tracking[15 – 16], and therefore we have found an opportunity that Wi-Fi has the feasibility of performing fall detection.

Most existing work can be divided into two stages: feature extraction and feature classification. Feature extraction refers to the extraction of parameters from the original Wi-Fi signal as human fall features, and feature classification refers to the construction of classifiers to identify fall actions based on various classification methods. In feature extraction, most existing work can be divided into two categories. The first category is to directly use the original features in the received signal, mainly the amplitude and phase information of CSI, including WiFall[17], RT-Fall[18], Anti-Fall[19], FallViewer[20], etc. WiFall is the first work that utilizes commercial Wi-Fi devices for fall detection, which characterizes human activity by using the fluctuation of the amplitude of CSI. Furthermore, Anti-fall combines the amplitude and phase information of CSI to characterize human activities. Considering that the phase of CSI collected by commercial devices contains random noise, RT-Fall reduces the impact of random phase by calculating the phase difference of different receiving antennas and uses the amplitude and phase difference information of CSI to demonstrate human activities. FallViewer designs a series of CSI denoising schemes to obtain more refined CSI information for characterizing human activities. These works all utilize traditional machine learning approaches such as support vector machines (SVM) to classify activity and perform fall detection. The second category is to process the received signal in the time and frequency domain to obtain higher-order features, such as the short time Fourier transform (STFT) spectrum, and this type of work includes FallDeFi[21], TL-Fall[22], CNNFall[23], etc. Both FallDeFi and CNNFall utilize denoising of CSI and perform STFT to extract Doppler frequency shift information to characterize user activities. FallDeFi extracts statistical features from the Doppler frequency shift and selects the features that are closely related to human activities, using SVM for classification. On the other hand, CNNFall uses a convolutional neural network model to classify Doppler information and perform fall detection. The original features of CSI are influenced by the environment, which makes it difficult to represent human activities; the STFT spectrum has a fixed window size, which cannot balance the temporal and spatial resolution. Most existing works[17 – 22] in feature classification have used traditional machine learning schemes for classification, such as SVM and random forests. A small number of works have implemented a classification based on deep learning models[23]. In these traditional machine learning solutions, statistical features are extracted without clear physical meanings. In order to better meet the practical needs, we are devoted to designing a fall detection system based on commercial Wi-Fi devices, which increases the system performance and minimizes the computational complexity of the system. In the feature extraction stage, we extract the discrete wavelet transform (DWT) spectrum from the received raw signal as a feature to characterize the person ' s activities. Theoretically, using different window sizes to extract human activity information in different frequency bands of the DWT spectrum can maintain high frequency resolution in the low-frequency band and high time resolution in the high-frequency band, which is more flexible than using a fixed window size. In addition, DWT spectrums can reduce the interference of the surrounding environment on the channel state information and mitigate the impact of environmental changes. In the feature classification stage, we design a deep learning model based on convolutional neural networks to build a classifier to identify the fall action. Compared with traditional machine learning models, deep models can better extract high-order spatio-temporal information about human activities and obtain more universal representations of human activities. Meanwhile, we evaluate the system based on the data collected on commercial Wi-Fi devices. The evaluation results show that the performance is better than existing fall detection work and other combined schemes, and the complexity of the model is less than other deep classification models. In summary, the main contributions of this paper are as follows.

1) We propose a new fall detection solution based on commercial Wi-Fi devices, which achieves better performance with less computation costs compared with existing solutions.

2) We select the discrete transform wavelet spectrum as the feature for activity classification, which has better environmental robustness compared with the original feature.

3) We design a deep learning model based on convolutional neural networks, which can extract higher-order features and better characterize human activities compared with traditional machine learning models.

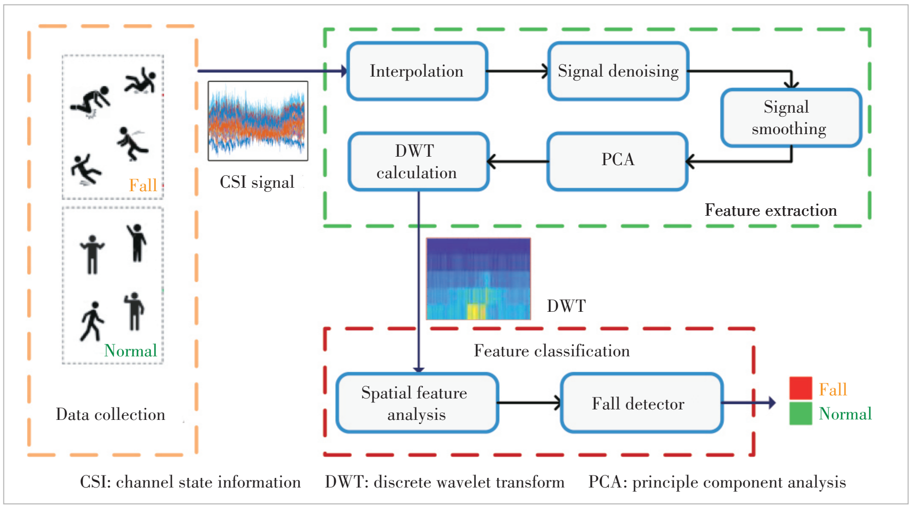
4) We have conducted extensive experiments, including both fall and non-fall. The experimental results show that our work outperforms existing fall detection work and other combined schemes, and achieves a false alarm rate of 4.8% and a missed alarming rate of 1.9%.

The rest of this work is organized as follows. Section 2 presents the design of the system, Section 3 evaluates the implemented system, and Section 4 summarizes this work.

2 System Design

2.1 System Overview

The overall framework design of the system is shown in Fig. 1. In the data collection module, we use commercial Wi-Fi devices to collect CSI of different activities of people. The feature extraction module preprocesses the received raw CSI information and then extracts the DWT spectrum as the features for activity classification. The feature classification



▲Figure 1. System framework diagram

module uses a deep learning model to analyze the spatial features of the input feature spectrum images and perform binary classification to determine the presence of fall activities. We will introduce feature extraction and feature classification respectively in Sections 2.2 and 2.3.

2.2 Feature Extraction

The feature extraction module extracts the corresponding features from the CSI information received by the commercial Wi-Fi device, which is used to characterize the activity of a person to identify the person's activity. CSI reflects the information of the physical layer channel and represents the channel response of the wireless link[24]. CSI is the channel attribute of the communication link, which represents the fading factor of the signal between the transmitter and the receiver for each transmission path between the transmitter and the receiver as a fading factor. Let $X(f,t)$ and $Y(f,t)$ be the frequency domain responses of the transmitter and receiver at moment t and subcarrier f, respectively, then the following relationship exists between them,

$$Y(f,t) = H(f,t) \times X(f,t), \tag{1}$$

where $Y(f,t)$ represents the channel frequency response (CFR), which is the frequency domain representation of CSI, and is usually a complex value. In practice, there are usually multiple propagation paths between the transmitter and the receiver, so it can be written in the following form:

$$H(f,t) = \sum_{k=1}^{N} \alpha_k(f,t) e^{-j2\pi f \tau_k(t)}, \tag{2}$$

where N is the number of multipaths, and $\alpha_k(f,t)$ and $\tau_k(t)$ represent the attenuation coefficient of the k-th propagation path and the propagation delay, respectively. In this experiment, the CSI information obtained from each receiver antenna contains 30 subcarriers[25]. In this system, we use the amplitude information of CSI for subsequent data processing.

As shown in Fig. 1, feature extraction in this system mainly includes signal interpolation, signal denoising, signal smoothing, principal component analysis, and DWT calculation[22].

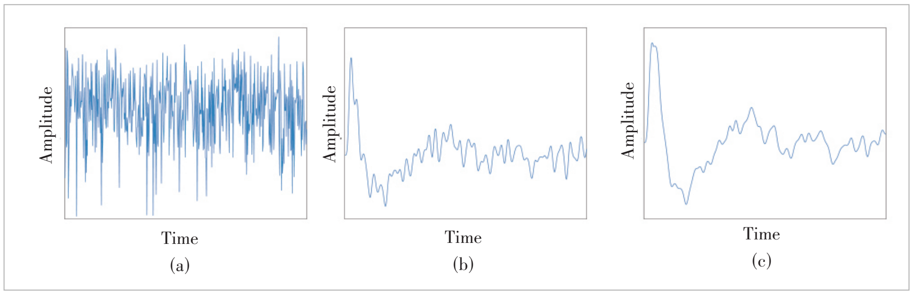
1) The purpose of signal interpolation is to obtain uniformly distributed samples. During the transmission of Wi-Fi signals, due to airport blocking and other reasons, the received data packets may have uneven sampling in the time domain. The theoretical analysis of the time-frequency domain in signal processing is based on the assumption of uniform sampling. Therefore, if the actual sampling is non-uniform, the results of video analysis will contain noise, and interference frequencies that are not present in the original signal will appear in the spectrum. This can make the extracted time-frequency domain features unable to fully reflect the activity information of the person, which will affect the classification of the person's activity. In the system, we perform one-dimensional linear interpolation on the CSI amplitude information extracted from non-uniform sampling[18] to reduce the impact of non-uniform sampling.

2) In the signal denoising part, the original CSI is filtered to retain the main components of personnel activities and filter out high-frequency and low-frequency noise. As shown in Fig. 2(a), the original CSI signal usually contains a lot of noise. The noise includes not only low frequency noise such as hardware noise and DC components but also high frequency noise such as signal burst. The main component of human activities that we need is in the middle of high frequency and low frequency. Therefore, we use a band-pass Butterworth filter for filtering. The setting of the low cutoff frequency is based on a balance between the requirements of interference elimination and the loss of low-frequency information. Specifically, the speed of normal human motion does not exceed 4 m/s. The Wi-Fi device operates at 5.825 GHz, and the corresponding Doppler frequency spectrum (DFS) upper limit is calculated to be 80 Hz[32]. Generally, the range of signal low-frequency noise is 0 Hz – 4 Hz, and the Doppler frequency deviation caused by personnel activities including falling, walking, bending and sitting is usually not more than 80 Hz. Therefore, we first carry out band-pass filtering on the signal, with a passband range of 4 Hz – 80 Hz, to filter out band noise[22]. The filtered CSI signal is shown in Fig. 2(b), and most of the disturbances in the signal have been filtered.

3) Signal smoothing is to better reduce the influence of in-band noise and signal jitter on activity recognition. We use the weighted moving average method to smooth the filtered CSI signal. Let's assume that the sampling sequence of a subcarrier of the CSI at different times is $C = [v_1, v_2, \cdots, v_L]$, then the smoothed CSI sequence is the weighted average of the CSI sampling values at the previous time, that is:

$$\hat{v}_i = \frac{1}{n + (n-1) + \cdots + 1} \times \sum_{j=1}^{n} (j \times v_{i-n+j}), \tag{3}$$

where $\hat{v}_i$ is the i-th sampling point of the smoothed CSI, and

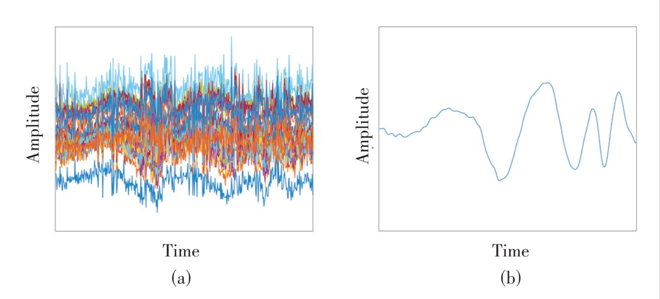


▲ **Figure 2. (a) Original channel state information (CSI) amplitude image; (b) CSI amplitude image after signal denoising; (c) CSI amplitude image after signal smoothing** the value of n represents the correlation between the CSI smoothing result and the CSI sampling at the past time. In the experiment, we take n=20. The smoothed CSI image is shown in Fig. 2(c).

4) The purpose of principal component analysis (PCA) is to extract the main features in the subcarrier to achieve more accurate activity recognition. In this work, each receiving antenna can obtain data from 30 subcarriers. We apply PCA to each subcarrier of CSI and select the second principal component of the signal for subsequent feature extraction, because the first principal component in the signal usually contains a lot of noise while containing the information about human activities. Fig. 3(a) shows the original amplitude of each CSI subcarrier, and Fig. 3(b) shows the second principal component of the corresponding CSI signal.

5) The purpose of DWT calculation is to obtain a discrete wavelet transform spectrum for fall detection. In the process of falling, people first have an acceleration process, and the acceleration is downward. The speed reaches the maximum when it collides with the ground quickly, and then the ground gives people an upward force. The acceleration is upward, and the speed quickly drops to 0. Compared with STFT[26], DWT can achieve a good trade-off between time resolution and frequency domain resolution. In the higher frequency range, actions usually change quickly, which can achieve higher time resolution; in the lower frequency range, the action usually changes slowly and can achieve higher frequency domain resolution. In order to accurately detect the change in the user's motion speed, in this work, we use the time-frequency domain component of the signal to detect falls[27]. DWT can calculate the corresponding energy size of components in different frequency ranges[28 - 29].

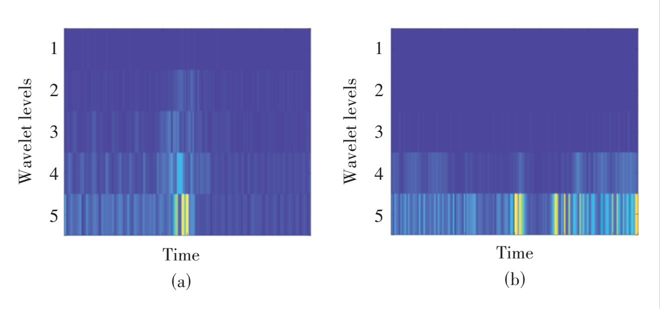
In this work, we use the demy wavelet base to obtain the fifth-order DWT spectrum. In the experiments, we find that the demy wavelet basis with five levels of frequency order is more suitable for fall detection. Typically, fall actions bring higher signal frequencies, with the highest values usually at level 4 or level 5, while non-fall actions usually have frequency orders below level 3. At the same time, using a moderate number of levels also reduces the computational burden of subsequent calculations. Fig. 4 shows the DWT spectrum of fall and walk activities. The place with higher brightness represents higher signal energy. It can be seen that the energy of signals in fall activities gradually increases from level 5 to level 2, and then gradually decreases. The energy of the signal in the walking motion is always at a lower wavelet level. It can be seen that the wavelet energy distribution of signals varies with different activity types.



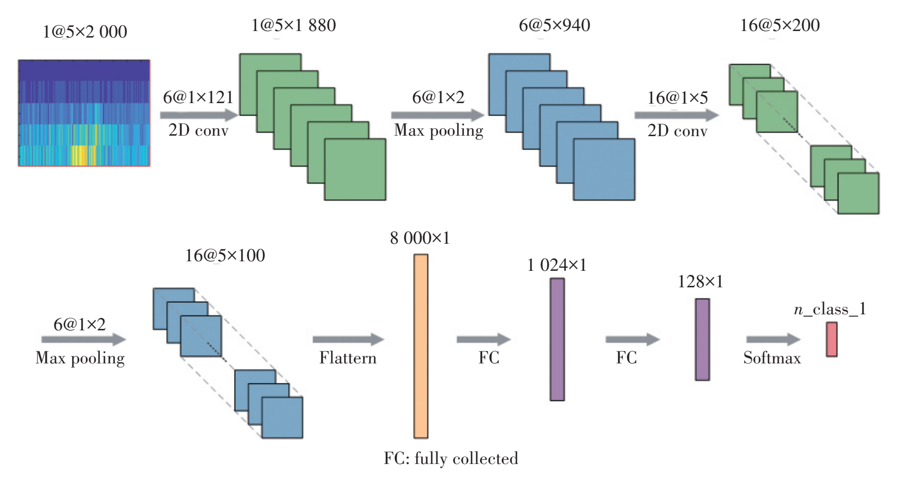
▲**Figure 3. (a) Amplitude image of each subcarrier of the original channel state information (CSI) and (b) second principal component amplitude image**

2.3 Feature Classification

The DWT spectrum can reflect the time-frequency domain characteristics of the signal, and this part uses convolutional neural networks (CNN) to classify the extracted DWT spectrum as shown in Fig. 5. In theory, the horizontal axis of the DWT spectrum we extract represents the time information of the user's activity, while the vertical axis represents the frequency information of the user's activity. CNN uses convolutional kernels of different sizes to extract edge information from the DWT spectrum. The horizontal component of the convolutional kernel can extract the temporal difference in-



▲**Figure 4. (a) Discrete wavelet transform (DWT) spectrum of fall activities and (b) DWT spectrum of walking motion**



▲**Figure 5. Feature classification model** formation of the user's activity, while the vertical component can extract the frequency difference information of the user's activity. Overall, using a CNN model can extract high-order information about the user's activity in both time and space domains, thus obtaining a high-order representation of the user's activity.

Let S be the input data set. This work pre-collects sample data of falls as well as normal activities and extracts features for training, where each sample datum has a time length of T = 2 s and a sampling frequency of 1 000 Hz. Each input sample of the classification model $s \in S$ is a DWT 2D spectrum of 5×2 000, where the sampling length of the time dimension is 2 000 and the frequency dimension is quantified into 5 levels.

Our work first extracts the spatial features of the 2D spectrum using convolutional and pooling layers[30 – 31].

$$F = g_2\Big(f_2\Big(g_1\Big(f_1\big(S, \theta_1\big), \theta_2\Big), \theta_3\Big), \theta_4\Big), \tag{4}$$

where f_1 and f_2 represent the convolutional layers, g_1 and g_2 represent the pooling layers, $\theta_1 - \theta_4$ represent the parameters, and F represents the extracted spatial features. Specifically, we first generate six feature maps of dimension 5×1 880 using six convolutional kernels of dimension 1×121, and then generate six feature maps of dimension 5×940 using the maximum pooling layer. Then we continue to generate 16 feature maps of dimension 5×200 using 16 convolutional kernels of dimension 1×5, and then generate 16 feature maps of dimension 5×100 using the maximum pooling layer. The feature maps are then generated using the maximum pooling layer. With two convolutional and pooling layers, we extract the spatial features of the signal. Next, we spread the dimensionality of the features and input them to the subsequent fully connected (FC) and Softmax layers for fall detection.

$$R = \text{softmax}\Big(h_2\Big(h_1\big(F, \theta_5\big), \theta_6\Big), \theta_7\Big), \tag{5}$$

where h_1 and h_2 represent the fully connected layer and $\theta_5 - \theta_7$ represent the parameters. The FC layers are activated using rectified linear units (ReLU) and each FC layer uses a dropout mechanism to avoid overfitting. In this way, we use features to determine the presence of dropout activity.

The system's overall algorithm is shown in Algorithm 1.

Algorithm 1. Fall detection algorithm

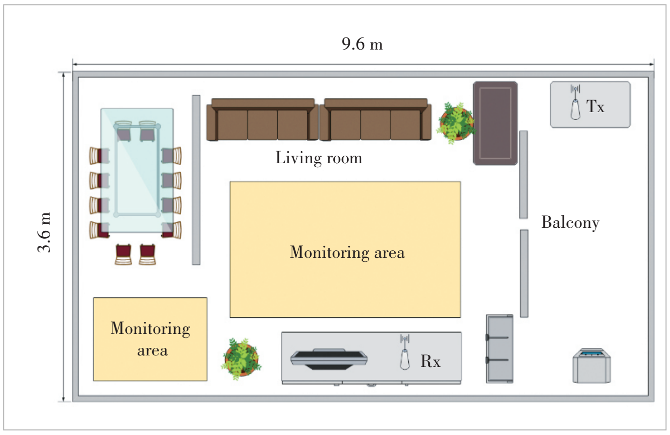
```
Input: CSI$_{raw}$, the raw CSI measurements.
Output: Fall detection results.
Signal interpolation:
CSIinterp ⟵ CSIraw
Signal denoise:
CSIdenoise ⟵ CSIinterp
Signal smoothing:
CSIsmooth ⟵ CSIdenoise
Principal component analysis:
PCs ⟵ PCA (CSI$_{smooth}$) PC$_2$ ⟵ the second PCs calculate DWT Spectrum:
DWT spectrum ⟵ DWT (PC$_2$) with demy wavelet base classify falls and non-falls:
Fall detection results ⟵ deep model (trained model, DWT Spectrum)
return Fall detection results
```

3 System Evaluation

3.1 Experiment Methodology

1) Experimental setup. The goal of our work is to implement a low-cost, senseless, non-contact fall detection system, so this paper uses a commercially available Wi-Fi device for the experiments. The experiments are based on a previously acquired dataset, the acquisition environment of which is shown in Fig. 6, and the size of the common home environment is 9.6 m×3.6 m. The yellow area is the fall monitoring area, and the line-of-sight path between the transmitter and receiver is obscured by a door. We use the CSITools platform and an Intel 5300 wireless card to collect CSI information. The center frequency of the wireless cards for both the transmitter and receiver is set to 5.825 GHz with a bandwidth of 20 MHz. The receiver is set to monitor modes to receive data from the transmitter. The transmitter sends CSI information at a frequency of 1 000 Hz.

2) Data acquisition. Our work collects data in the monitoring area in Fig. 6. There are five members in the family of the experiment. To obtain data on falls, this work asks participants to perform the fall action on their own with controlled risk. Specifically, participants wear protective equipment and pretend to fall unconsciously whenever possible. To obtain more data on falls, we also use dummies to simulate real users to perform falls. In addition, we collect non-falling activities of each user in their daily life. To collect data on normal activities, users are asked to perform activities in the monitored area. In total, about 600 sets of fall samples and 2 000 sets of non-fall samples are collected in this work. Among the fall



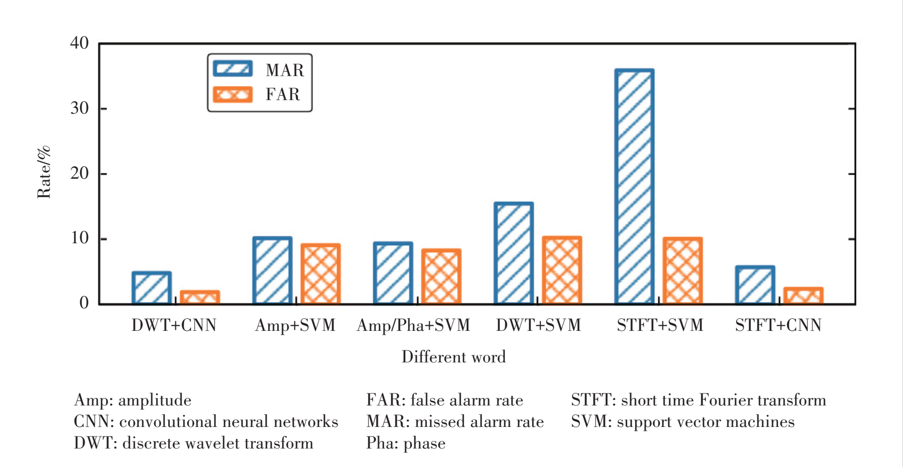
▲Figure 6. Experimental environment setup samples, the number of dummy samples accounts for about 490 groups, and the rest are falls of real users. The sample types of falls include tripping, slipping, losing balance, kneeling, sitting-falling, and walking-falling, and the sample types of non-falls include activities such as walking, jogging, sitting/standing up, bending down to pick up, and squatting.

3) Detection metrics. Our work uses two intuitive fall detection metrics: the false alarm rate (FAR) and the missed alarm rate (MAR). FAR is the ratio between the number of incorrectly identified normal activity samples and that of all normal samples, showing how often users are disturbed when no fall activity occurs. MAR is the ratio between the number of incorrectly identified fall samples and that of all fall samples, showing the sensitivity and detection capability of the system for fall activity.

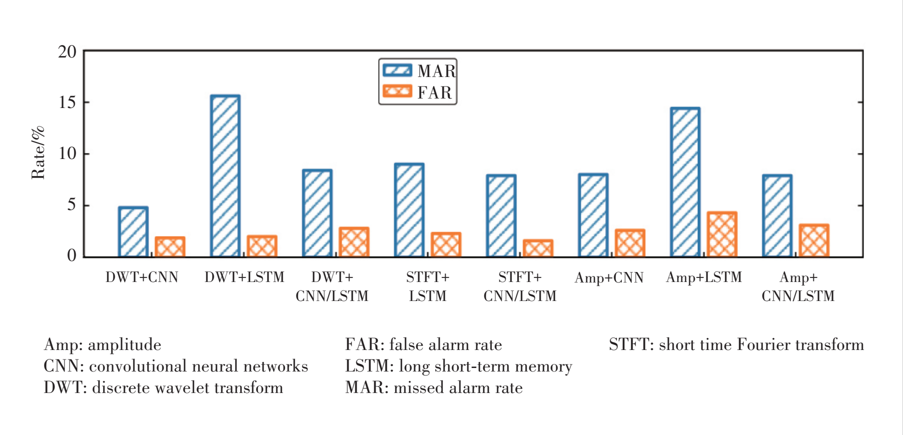
3.2 System Performance

1) Performance comparison of existing work. We compare the present work with the currently available work on fall detection using Wi-Fi based devices. We divide the pre-collected dataset into a training set and a test set and extract the corresponding features for evaluation, using a ten-fold cross-validation approach. Fig. 7 shows the performance of the system evaluation. The MAR and FAR of our system are 4.8% and 1.9%, respectively, which are better than the existing work. Since the experimentally collected non-fall data are all data of users performing activities, and according to the survey results of the National Bureau of Statistics, the average time spent by Chinese residents at home is about 7.5 h per day, and the FAR of the system will be further reduced in the home scenario, which is expected to be around 0.6%.

A comparison of existing work shows that using deep models for fall detection performs significantly better than using traditional machine learning. For example, using a deep learning model outperforms a traditional machine learning SVM model when the same DWT spectrum is used as the extracted feature. Specifically, the MAR and FAR decrease by 10.6% and 8.3%, respectively. In theory, the deep learning model can acquire more hidden features in the wireless signal; while the traditional machine learning model mostly extracts statistical features for activity identification, which is relatively less physically significant.



▲Figure 7. Comparison of existing work



▲Figure 8. Comparison of depth modeling solutions

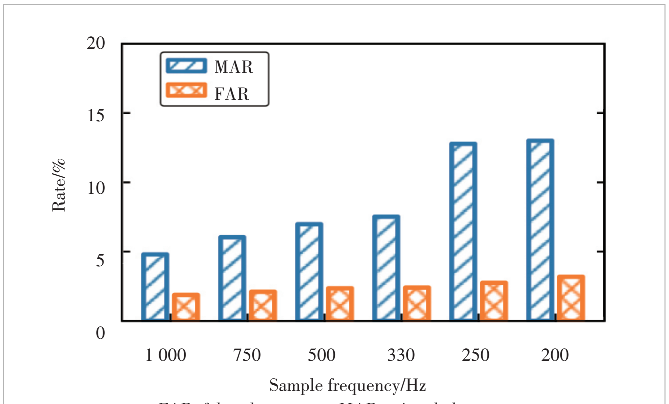
2) Performance comparison of different deep model schemes. In addition to the already working detection schemes, we combine different detection schemes by ourselves based on the CSI amplitude/phase, DWT spectrum and STFT spectrum, combined with models such as the long short-term memory (LSTM) network in deep learning. The results of our systematic evaluation of different schemes are shown in Fig. 8. It can be seen that with the deep learning models, the performance of the DWT spectrum and STFT spectrum-based schemes is better than that of the original feature-based schemes such as CSI magnitude. For example, when the CNN model is used for feature classification, the DWT spectrum is better than the CSI amplitude. Specifically, MAR and FAR decrease by 3.2% and 0.7% respectively.

3) Analysis of CSI sampling frequency. In the above experiments, the CSI sampling frequency is 1 000 Hz. Considering that the Wi-Fi packet transmission may be disturbed in actual use, the sampling frequency may be degraded. Therefore, we evaluate the impact of the CSI sampling frequency on the system performance. We use 1 000 Hz to capture CSI, downsample the CSI data stream to 750 Hz, 500 Hz, 330 Hz, 250 Hz

and 200 Hz, and adjust the input scale of the network to match the extracted signal features for fall detection. The performance of the system is shown in Fig. 9, where the performance of the system also decreases gradually as the sampling frequency decreases. When the sampling frequency decreases from 1 000 Hz to 200 Hz, the MAR and FAR of the system decrease by 8.2% and 1.3%, respectively. This is due to the fact that the user's velocity increases suddenly during the falling motion, which is harder to capture at lower sampling frequencies.

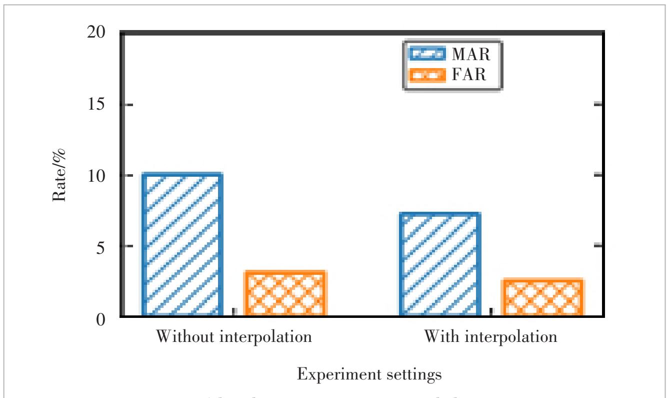
4) Signal interpolation algorithm performance analysis. To test the performance of the signal interpolation algorithm in the case of non-uniform sampling, we construct non-uniformly sampled data by randomly selecting 50% of the existing uniformly sampled samples with an overall sampling frequency of 500 Hz, and process the constructed data using a one-dimensional linear interpolation method. We then perform feature extraction and classification on the constructed non-uniformly sampled data and the interpolated data respectively, and observe the effect. As shown in Fig. 10, after signal interpolation, the MAR and FAR of the system are 7.2% and 2.5%, respectively. The results are 2.8% and 0.6% lower than the MAR and FAR of the directly non-uniformly sampled data, respectively. Theoretically, using non-uniformly sampled data to calculate the time-frequency domain characteristics of the signal introduces a certain amount of error. The interpolation algorithm of the signal can mitigate this part of the error.

5) Relevant parameters and activity analysis. In the experiment, the MAR and FAR of the system are also changed by adjusting the threshold of activity discrimination. Fig. 11 shows the changes in MAR and FAR in the case of system threshold adjustment. It can be seen that MAR and FAR constrain each other, and theoretically, the thresholds can be adjusted as needed to obtain the corresponding performance of the system. In this system, we adjust the MAR around 4.8% and obtain the corresponding FAR of 1.9%. In addition, we analyze the probability of misjudgment for different normal activities and the probability of misjudgment for different fall types, and the results are shown in Figs. 12 and 13, respectively. It can be seen that the bending and picking up action has the highest false alarm rate of 4.0%, followed by walking and sitting/standing up. The speed of human movements in these actions is usually faster, and the actions of bending down and picking up, and sitting down/standing up have some similarities with falls, so false alarms occur easily. And among the different types of falls, tripping has the highest missed alarm rate of 7.8%, followed by kneeling and sitting/stumbling. Since tripping and kneeling happen when the user usu-



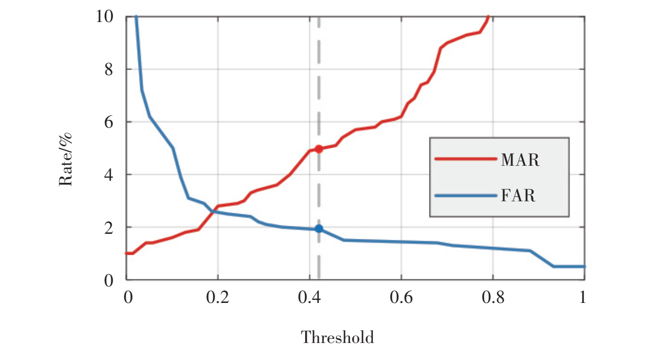
FAR: false alarm rate MAR: missed alarm rate

▲Figure 9. Effect of different channel state information (CSI) sampling frequencies



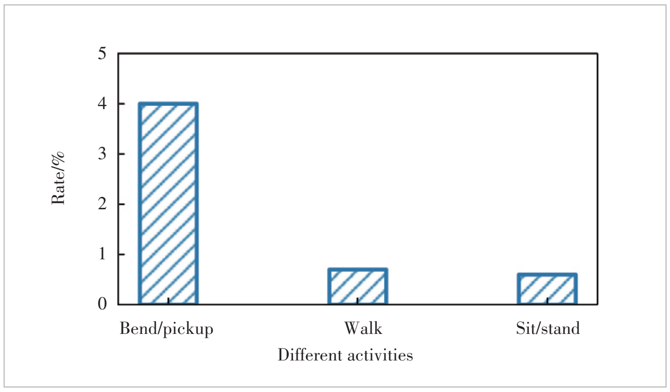
FAR: false alarm rate MAR: missed alarm rate

▲Figure 10. Signal interpolation algorithm performance analysis

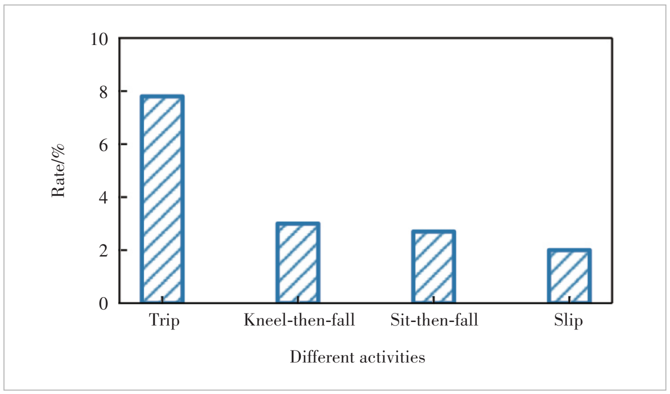


FAR: false alarm rate MAR: missed alarm rate

▲Figure 11. Impact of threshold selection on performanc



▲Figure 12. Probability of false alarms for different non-falling activities

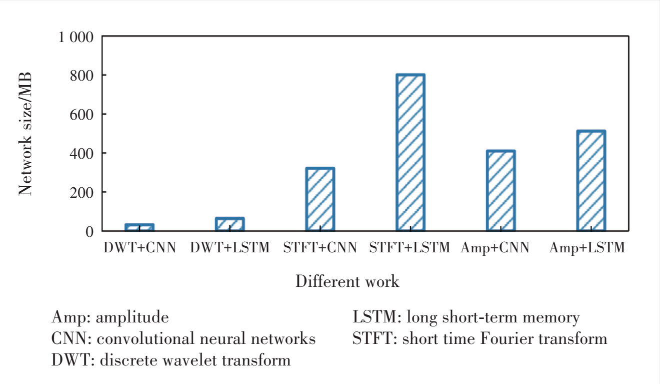


▲**Figure 13. Probability of missing alarms for different fall types**

ally falls toward the front of the direction of motion, there is a certain similarity with the action of bending down to pick up, while the sit-down-fall situation is easily confused with sitting down and therefore easily to miss.

6) Comparison of network sizes for different deep models. We compare the scale of various deep networks, and the results of the comparison are shown in Fig. 14. In the case of different features, the corresponding network models are smaller in size because the DWT spectrum has fewer orders of features compared with the original CSI magnitude information and the STFT spectrum. From the results, the scale of the model corresponding to using the DWT spectrum as features is one order of magnitude less than the other two features. For the different classification models, the CNN model has fewer parameters than the LSTM model with the same size input, so the network size is smaller. Overall, our system achieves better performance by using a network model as small as possible.

7) System latency analysis. To validate the efficiency of our system, we deployed it on a laptop with an 8-core Intel i7-6700 @2.60 GHz CPU and measured the system's runtime. The system's runtime is mainly composed of feature extraction and model classification. Experimental results show that using 2 s of CSI information as input, the overall average end-to-end runtime is 18.1 ms, with feature extraction taking 7.5 ms and model classification taking 10.6 ms. The results indicate that our system can achieve real-time detection of fall actions.



▲**Figure 14. Network size for models with different depth**

4 Conclusions

In our work, a passive fall detection system based on Wi-Fi is proposed. To better obtain information about the motion state of the target, this work extracts the DWT spectrum from the received raw signal to characterize the user's activity. To achieve better classification results, this work designs a classifier based on a deep learning model for fall detection. The experimental evaluation illustrates that our work achieves false alarm and missed alarm rates of 4.8% and 1.9%, with better performance than other existing works and systems.

References

[1] EROL B, AMIN M G, BOASHASH B, et al. Wideband radar based fall motion detection for a generic elderly [C]//The 50th Asilomar Conference on Signals, Systems and Computers. IEEE, 2017: 1768 – 1772. DOI: 10.1109/ACSSC.2016.7869686

[2] BIAN Z P, HOU J H, CHAU L P, et al. Fall detection based on body part tracking using a depth camera [J]. IEEE journal of biomedical and health informatics, 2015, 19(2): 430 – 439. DOI: 10.1109/JBHI.2014.2319372

[3] STONE E E, SKUBIC M. Fall detection in homes of older adults using the microsoft kinect [J]. IEEE journal of biomedical and health informatics, 2015, 19 (1): 290 – 301. DOI: 10.1109/JBHI.2014.2312180

[4] KAU L J, CHEN C S. A smart phone-based pocket fall accident detection, positioning, and rescue system [J]. IEEE journal of biomedical and health informatics, 2015, 19(1): 44 – 56. DOI: 10.1109/JBHI.2014.2328593

[5] PIERLEONI P, BELLI A, PALMA L, et al. A high reliability wearable device for elderly fall detection [J]. IEEE sensors journal, 2015, 15(8): 4544 – 4553. DOI: 10.1109/JSEN.2015.2423562

[6] RAMEZANI R, XIAO Y B, NAEIM A. Sensing-Fi: Wi-Fi CSI and accelerometer fusion system for fall detection [C]//Proceedings of 2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI). IEEE, 2018: 402 – 405. DOI: 10.1109/BHI.2018.8333453

[7] FOROUGHI H, ASKI B S, POURREZA H. Intelligent video surveillance for monitoring fall detection of elderly in home environments [C]//The 11th International Conference on Computer and Information Technology. IEEE, 2009: 219 – 224. DOI: 10.1109/ICCITECHN.2008.4803020

[8] JUDERAJENDRAN P, DALALI S. A smart and passive floor-vibration based-fall detector for elderly [C]//The 2nd International Conference on Information & Communication Technologies. IC-TTA, 2006. DOI: 10.1109/ICTTA.2006.1684511

[9] HAN Y T, LI H, ZHU G X, et al. Indoor target detection and localization method based on WiFi [J]. ZTE technology journal. 2022, 27(5): 46 – 52. DOI: 10.12142/ZTETJ.202205009

[10] LI F L, YANG W C, ZHANG X B. Design and application on collaborative networking scheme of 5G and WiFi6 [J]. ZTE technology journal. 2022, 27(4): 7 – 13. DOI: 10.12142/ZTETJ.202204003

[11] ABDELNASSER H, YOUSSEF M, HARRAS K A. WiGest: A ubiquitous WiFi-based gesture recognition system [C]//IEEE Conference on Computer Communications (INFOCOM). IEEE, 2015: 1472 – 1480. DOI: 10.1109/INFOCOM.2015.7218525

[12] JIANG W J, MIAO C L, MA F L, et al. Towards environment independent device free human activity recognition [C]//The 24th Annual International Conference on Mobile Computing and Networking. ACM, 2018. DOI: 10.1145/3241539.3241548

[13] KORANY B, KARANAM C R, CAI H, et al. XModal-ID: Using WiFi for through-wall person identification from candidate video footage [C]//The 25th

Annual International Conference on Mobile Computing and Networking. ACM, 2019. DOI: 10.1145/3300061.3345437

[14] ZENG Y Z, PATHAK P H, MOHAPATRA P. WiWho: WiFi-based person identification in smart spaces [C]//The 15th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN). IEEE, 2016: 1 – 12. DOI: 10.1109/IPSN.2016.7460727

[15] LI X, ZHANG D Q, LV Q, et al. IndoTrack: Device-free indoor human tracking with commodity Wi-Fi [J]. Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies, 2017, 1(3): 1 – 22. DOI: 10.1145/3130940

[16] XIE Y X, XIONG J, LI M, et al. mD-track: Leveraging multi-dimensionality for passive indoor Wi-Fi tracking [C]//The 25th Annual International Conference on Mobile Computing and Networking. ACM, 2019: 1 – 16. DOI: 10.1145/3300061.3300133

[17] WANG Y X, WU K S, NI L M. WiFall: Device-free fall detection by wireless networks [J]. IEEE transactions on mobile computing, 2017, 16(2): 581 – 594. DOI: 10.1109/TMC.2016.2557792

[18] WANG H, ZHANG D Q, WANG Y S, et al. RT-fall: A real-time and contactless fall detection system with commodity WiFi devices [J]. IEEE transactions on mobile computing, 2017, 16(2): 511 – 526. DOI: 10.1109/TMC.2016.2557795

[19] ZHANG D Q, WANG H, WANG Y S, et al. Anti-fall: a non-intrusive and real-time fall detector leveraging CSI from commodity WiFi devices[C]//International Conference on Smart Homes and Health Telematics. Springer, 2015: 181 – 193.10.1007/978-3-319-19312-0_15

[20] WANG Y C, YANG S, LI F, et al. FallViewer: A fine-grained indoor fall detection system with ubiquitous Wi-Fi devices [J]. IEEE Internet of Things journal, 2021, 8(15): 12455 – 12466. DOI: 10.1109/JIOT.2021.3063531

[21] PALIPANA S, ROJAS D, AGRAWAL P, et al. FallDeFi: ubiquitous fall detection using commodity Wi-Fi devices [J]. Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies. ACM, 2019. DOI: 10.1145/3161183

[22] ZHANG L, WANG Z R, YANG L. Commercial Wi-Fi based fall detection with environment influence mitigation [C]//The 16th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON). IEEE, 2019: 1 – 9. DOI: 10.1109/SAHCN.2019.8824989

[23] NAKAMURA T, BOUAZIZI M, YAMAMOTO K, et al. Wi-Fi-CSI-based fall detection by spectrogram analysis with CNN [C]//IEEE Global Communications Conference. IEEE, 2021: 1 – 6. DOI: 10.1109/GLOBECOM42002.2020.9322323

[24] YANG Z, ZHOU Z M, LIU Y H. From RSSI to CSI: indoor localization via channel response [J]. ACM computing surveys, 46(2): 1 – 32. DOI: 10.1145/2543581.2543592

[25] XIAO Y. IEEE 802.11n: Enhancements for higher throughput in wireless LANs [J]. IEEE wireless communications, 2005, 12(6): 82 – 91. DOI: 10.1109/MWC.2005.1561948

[26] QIAN K, WU C S, ZHOU Z M, et al. Inferring motion direction using commodity Wi-Fi for interactive exergames [C]//The 2017 CHI Conference on Human Factors in Computing Systems. ACM, 2017: 1961 – 1972. DOI: 10.1145/3025453.3025678

[27] GRIFFIN D, LIM J. Signal estimation from modified short-time Fourier transform [J]. IEEE transactions on acoustics, speech, and signal processing, 1984, 32(2): 236 – 243. DOI: 10.1109/TASSP.1984.1164317

[28] WANG W, LIU A X, SHAHZAD M, et al. Understanding and modeling of WiFi signal based human activity recognition [C]//The 21st Annual International Conference on Mobile Computing and Networking. ACM, 2015: 65 – 76. DOI: 10.1145/2789168.2790093

[29] SHENSA M J. The discrete wavelet transform: wedding the Atrous and Mallat algorithms [J]. IEEE transactions on signal processing, 1992, 40(10): 2464 – 2482. DOI: 10.1109/78.157290

[30] NOBRE J, NEVES R F. Combining principal component analysis, discrete wavelet transform and XGBoost to trade in the financial markets [J]. Expert systems with applications, 2019, 125: 181 – 194. DOI: 10.1016/j.eswa.2019.01.083

[31] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks [J]. Communications of the ACM, 2017, 60(6): 84 – 90. DOI: 10.1145/3065386

[32] SHIN H C, ROTH H R, GAO M C, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning [J]. IEEE transactions on medical imaging, 2016, 35(5): 1285 – 1298. DOI: 10.1109/TMI.2016.2528162

Biographies

GONG Panyin received his bachelor's degree in software engineering from Huazhong University of Science and Technology, China in 2018 and is currently studying for a master's degree in the School of Software, Tsinghua University, China. His research interests include the Internet of Things and wireless sensing.

ZHANG Guidong received his BE degree from the Department of Electronic Engineering and Information Science, University of Science and Technology of China in 2018. He is currently working toward his PhD degree with the School of Software, Tsinghua University. His research interests include wireless sensing and mobile computing.

ZHANG Zhigang graduated from Xi'an Jiaotong University, China. Currently, he is the planning director of the cable FM product team of ZTE Corporation. With more than 10 years of experience in research and planning of telecommunication products, he has accumulated and practiced for many years in home networks, smart homes, IP protocols, etc. He has participated in translating and publishing the textbook, *Principles of Compilation*, applied for three patents, and made several special speeches in technical forums.

CHEN Xiao graduated from Nanjing University of Aeronautics and Astronautics, China. He is currently the director of the Wireline Architecture Department of ZTE Corporation. He has more than 20 years of experience in research and planning of telecommunication products and related technologies. He has organized many national science and technology projects, and published many papers in various publications. He is the leading inventor of many patents.

DING Xuan (dingx04@gmail.com) received his bachelor's degree from the School of Software, Tsinghua University, China in 2008, and PhD degree from the Department of Computer Science and Technology, Tsinghua University in 2014. He is currently a research assistant professor in the School of Software, Tsinghua University. His research interests include privacy-preserving computing, blockchain, RFID and wireless sensing.

Incident and Problem Ticket Clustering and Classification Using Deep Learning

FENG Hailin[1], HAN Jing[2], HUANG Leijun[1],
SHENG Ziwei[3], GONG Zican[2]

(1. Zhejiang A&F University, Hangzhou 310007, China;
2. ZTE Corporation, Shenzhen 518057, China;
3. Huazhong University of Science and Technology, Wuhan 430074, China)

DOI: 10.12142/ZTECOM.202304009

https: //kns.cnki.net/kcms/detail/34.1294.TN.20231115.1421.002.html, published online November 16, 2023

Manuscript received: 2022-12-19

Abstract: A holistic analysis of problem and incident tickets in a real production cloud service environment is presented in this paper. By extracting different bags of words, we use principal component analysis (PCA) to examine the clustering characteristics of these tickets. Then K-means and latent Dirichlet allocation (LDA) are applied to show the potential clusters within this Cloud environment. The second part of our study uses a pre-trained bidirectional encoder representation from transformers (BERT) model to classify the tickets, with the goal of predicting the optimal dispatching department for a given ticket. Experimental results show that due to the unique characteristics of ticket description, pre-processing with domain knowledge turns out to be critical in both clustering and classification. Our classification model yields 86% accuracy when predicting the target dispatching department.

Keywords: problem ticket; ticket clustering; ticket classification

Citation (Format 1): FENG H L, HANG J, HUANG L J, et al. Incident and problem ticket clustering and classification using deep learning [J]. ZTE Communications, 2023, 21(4): 69–77. DOI: 10.12142/ZTECOM.202304009
Citation (Format 2): H. L. Feng, J. Han, L. J. Huang, et al., "Incident and problem ticket clustering and classification using deep learning," *ZTE Communications*, vol. 21, no. 4, pp. 69–77, Dec. 2023. doi: 10.12142/ZTECOM.202304009.

1 Introduction

For cloud service providers, maintaining an outstanding service level agreement with minimum downtime and incident response time is critical to the business. In order to provide such a prominent high-level reliability and availability, IT operation plays an important role. However, the emergence of modern computing architectures, such as virtual machines, containers, server-less architecture, and micro-services, brings additional challenges to the management of such environments[1–2].

Problem and incident tickets have been a long-standing mechanism in carrying on any issues reflected by customers, or any alerts raised by monitoring systems. According to the Information Technology Infrastructure Library (ITIL) specification, the incident, problem, and change (IPC) systems fulfill the tracking, analysis, and mitigation of problems[3]. Change requests are nowadays mostly managed differently due to the practice of DevOps. Incident and problem tickets often share the same system and process. An incident or problem ticket usually starts with a short description of the problem that has been originally observed. The ticket itself may be augmented by the personnel assigned along the debugging and resolution process. There are also multiple software platforms and services to help enterprises manage those tickets, including BMC Remedy, IBM Smart Cloud Control Desk, SAP Solution Manager, ServiceNow, etc.[4]

However, dispatching an incident or problem ticket is still basically a manual process depending on human knowledge. Some of the ticket management systems offer insights such as agent skill level, capacity, and relevance. There are some early works attempting to dispatch tickets based on the agent’ s speed from historical data[5]. Our observation reveals that dispatching to individual agents might be a secondary issue. Instead, finding the matching department for a specific issue appears to be a primary one especially if a prompt resolution period is the desired outcome. It is not uncommon for some tickets to go through multiple departments before it lands on the right one. For example, a service unavailable problem might be caused by security settings, networking, hosting services, applications, or even databases, and the specific problem may be resolved by one of the departments or by multiple departments. Therefore, it is essential to find the most likely department, especially at the beginning when the problem was initially reported to resolve the issue efficiently. The specific technical challenge of classifying an early ticket is that the only available feature is problem description.

2 Related Work

Since the day when computer systems were created, IT operation has been a critical issue. With the prevalence of online services, in order to minimize system downtime and maintain premium service level agreements, IT operation plays a central role in achieving such a goal. Especially in today's highly distributed multi-layered cloud environment, it is untrivial to effectively find the matching departments to resolve the issue.

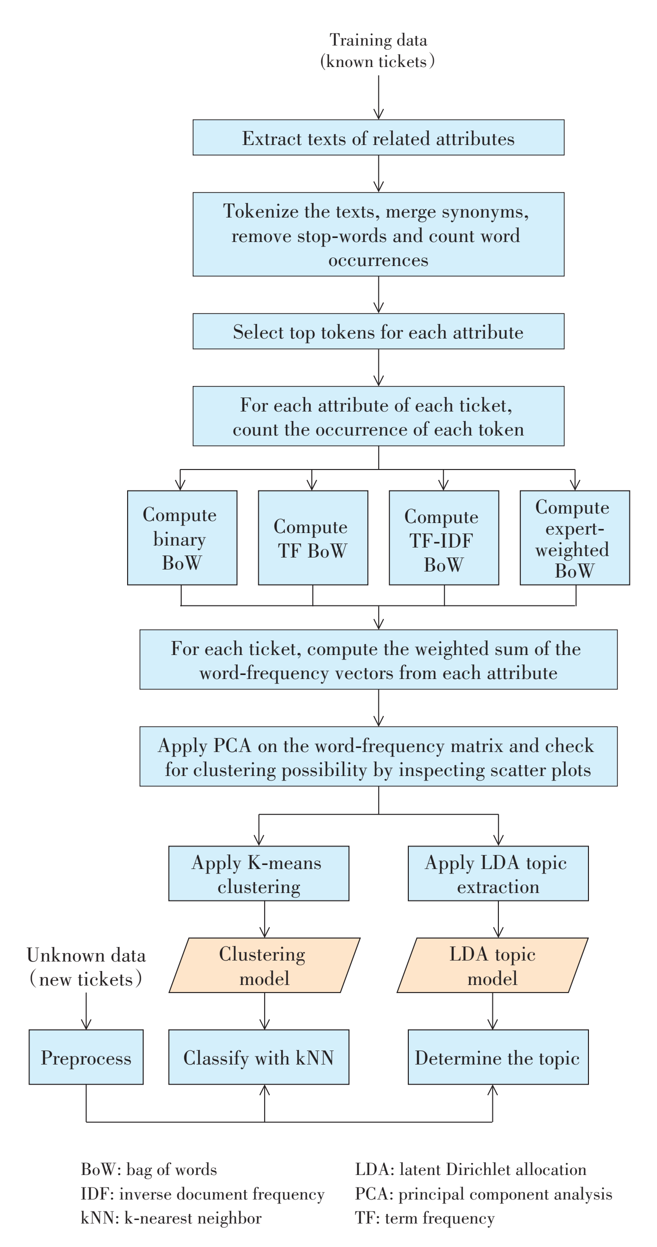
Artificial intelligence has been applied in IT operations, especially in anomaly detection[11–12], problem troubleshooting[13–14], and security[15–16]. A few works have attempted to improve the efficiency of ticket dispatching. BOTEZATU et al.[5] tried to find the most cost-effective agent for ticket resolution, rather than finding a matching group or department. SHAO et al.[17] focused on the transfer information in ticket resolution and formulated a model based on prior resolution steps. AGARWAL et al.[18] used a supported vector machine and a discriminative term to predict the matching department. While we use ticket descriptions and other attributes to find the best department, our solution is quite different from the previous works.

In terms of ticket analysis, there are only a few works on alerts or ticket clustering. LIN et al.[19] used graph theory and similarity measures as Jaccard as the cluster mechanism. MANI et al.[20] proposed a technique combining latent semantic indexing and a hierarchical *n*-gram algorithm. AGARWAL et al.[21] used a mixture of data mining, machine learning, and natural language parsing techniques to extract and analyze unstructured tests in IT tickets. JAN et al.[22] proposed a framework for text analysis in an IT service environment. We examine the clustering characteristics to discover the content of the ticket descriptions specific to the system under investigation. Our approach is generic to all systems with minor adjustments of synonyms and user dictionaries. When it comes to clustering itself, we believe our dataset is also unique as it is from the latest container-based cloud environment which is more complicated than prior systems.

3 Design of Clustering System

We apply different topic modeling algorithms to cluster the tickets based on their descriptions and compare their performance by calculating their sum of square error (SSE) and silhouette scores. The clustering results indicate the number of major topics in the ticket description corpus. Since it is an unsupervised learning process, it saves great effort from data annotation. For ticket classification, word embedding models have shown much better performance. Therefore, we only adopt the supervised approach using a pre-trained BERT model[6] which is fine-tuned with domain-specific labeled data.

Fig. 1 illustrates the overall steps we perform ticket description clustering. First, data preprocessing is performed by extracting texts, merging synonyms, removing stop words, etc. After tokenizing the texts, we construct 4 types of bags of words (BoW), including binary BoW, term frequency (TF) BoW, term frequency inverse document frequency (TF-IDF)[7] BoW, and expert-weighted BoW. For each of the BoW, we apply principal component analysis (PCA) to check for clustering possibility and use K-Means to cluster the topics. We also perform latent dirichlet allocation (LDA)[8] for topic extraction and modeling.


Training data
(known tickets)
Extract texts of related attributes
Tokenize the texts, merge synonyms, remove stop-words and count word occurrences
Select top tokens for each attribute
For each attribute of each ticket, count the occurrence of each token
Compute binary BoW
Compute TF BoW
Compute TF-IDF BoW
Compute expert-weighted BoW
For each ticket, compute the weighted sum of the word-frequency vectors from each attribute
Apply PCA on the word-frequency matrix and check for clustering possibility by inspecting scatter plots
Apply K-means clustering
Apply LDA topic extraction
Unknown data
(new tickets)
Clustering model
LDA topic model
Preprocess
Classify with kNN
Determine the topic
BoW: bag of words
IDF: inverse document frequency
kNN: k-nearest neighbor
LDA: latent Dirichlet allocation
PCA: principal component analysis
TF: term frequency

▲**Figure 1. Data analysis flow**

Finally, we show some of the sample topics in the cluster.

4 Experiments

We use two datasets from an enterprise-scale cloud provider, comprising 468 infrastructure-level and 787 Platform as a Service (PaaS) -level incident tickets, respectively. Since both datasets have similar data formats, we use the same analysis methods, which are mainly unsupervised machine learning approaches such as K-means and LDA. Our goal is to learn and make use of the inherent homogeneity of the complicated ticket descriptions by analyzing them.

For model training, we use the number, title or subject, and description from the datasets, in which the title or subject is a summary of the incident, and the description is a detailed text describing the problem. Some of the description texts are in the semi-structured form. For example, more than half the infrastructure-level ticket descriptions consist of explicit attributes like symptoms, progress, network topology, conclusions, steps, and remarks. We focus on the symptom attribute rather than using the entire text body since prediction needs to be performed when the ticket only has a symptom description. Some of the corpus such as file names, URL links, and system logs are filtered as part of preprocessing.

4.1 Data Preprocessing

We extract the text of the symptom attribute from the ticket description. If the description does not contain an explicit attribute of "symptom", the whole text is used. For the symptom text, we utilize regular expressions to filter unwanted data like picture-attached file name, date, time, URL and also delete the system logs as many as possible. We also perform spell checking using a dictionary.

Our next step is to convert the symptom texts into individual word tokens. Since most of the incident descriptions are a mixture of both Chinese and English, we use different tokenization tools for each language. "Jieba" is used for Chinese and "spaCy" for English. We also remove stop words from the output token and merge synonyms, e.g., "db" and database are the same, so they are uniformly replaced by a database. The lists of stop words are from Baidu[9] and github[10]. We merge both and extend some ticket-specific stop words for the experiments.

Given that some titles are similar to the symptom in terms of interfering texts and marks, they are preprocessed in the same way. The process described above ultimately generates a list of most frequently used tokens in both the title and symptom token lists. We sample high frequency Chinese and English words from the results, which are shown in Table 1.

4.2 Clustering Using BoW Models

First, we study the clustering characteristics of the incident tickets using the BoW model. For the tokens we extract during preprocessing, we choose the top high frequency words for title and symptom respectively. We combine the tokens from title and symptom based on a predefined weight so that each ticket is transformed into a word frequency vector, and accordingly, the dataset is represented by a word frequency matrix.

▼Table 1. Sample of high frequency words

Keywords	Frequency	Keywords	Frequency
node	538	tecs	328
defect	479	provider	198
version	463	daisy	150
symptom	329	nova	149
upgrade	317	dvs	139
alert	309	compute	123
conclusion	291	cinder	90
description	282	neutron	90
progress	275	sdn	79
operation	258	error	76
topology	256	host	69
note	253	nfv	69
cause	252	ip	64
site	235	agent	64
failure	229	ceph	62

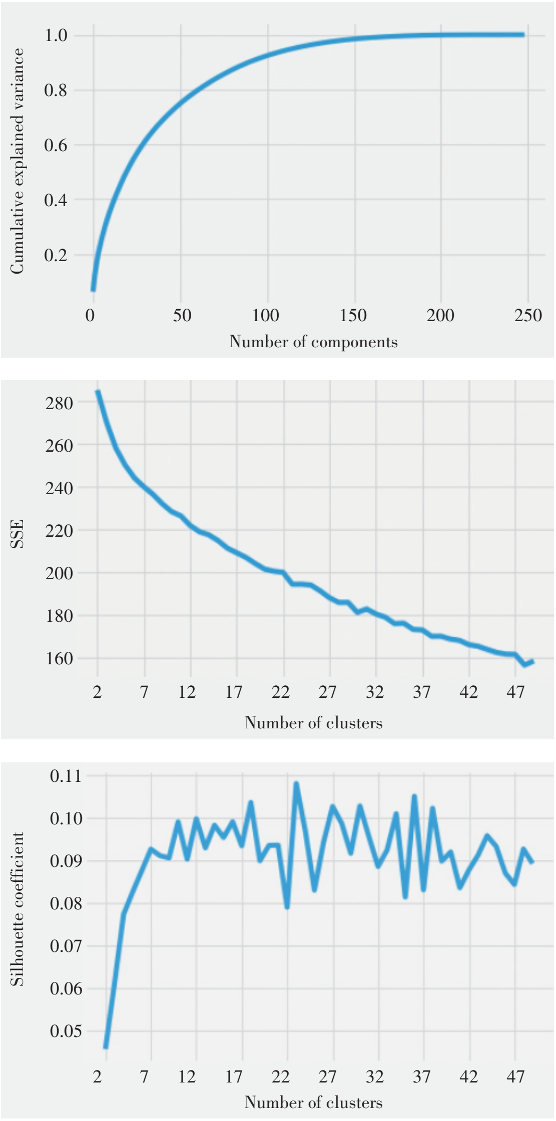
We apply principal component analysis (PCA) to the normalized word frequency matrix of the dataset, aiming to select the number of appropriate components using cumulative explained variance results. For example, Fig. 2 shows the PCA results of the symptom word frequency matrix, indicating that if 100 components are selected, and more than 90% of the variance can be explained.

After the number of the principal components is selected, we project the word frequency matrix to these components and use K-means for clustering analysis. For a given range of cluster numbers (i.e., values of K), we generate SSE and silhouette coefficient curves. As the best practice, the number of clusters is determined at the inflection point of the downward trend SSE curve or at the point when the upward trend silhouette coefficient curve becomes a plateau. The results are shown in Fig. 2. The SSE curve does not show an obvious inflection point, and the absolute value of the silhouette coefficient is too small even though the trend meets the demand (silhouette coefficient is between −1 and 1. The closer it is to 1, the more reasonable the clustering is). We conclude that it may not be a viable approach to evaluating the best cluster size by using PCA.

We also perform experiments using other models such as TF-IDF to generate a word frequency matrix, and the results are similar to PCA, indicating the word frequency matrix may not apply to incident tickets.

4.3 Clustering Using Latent Dirichlet Allocation (LDA) Model

In this section, we use LDA to extract dominant topics from

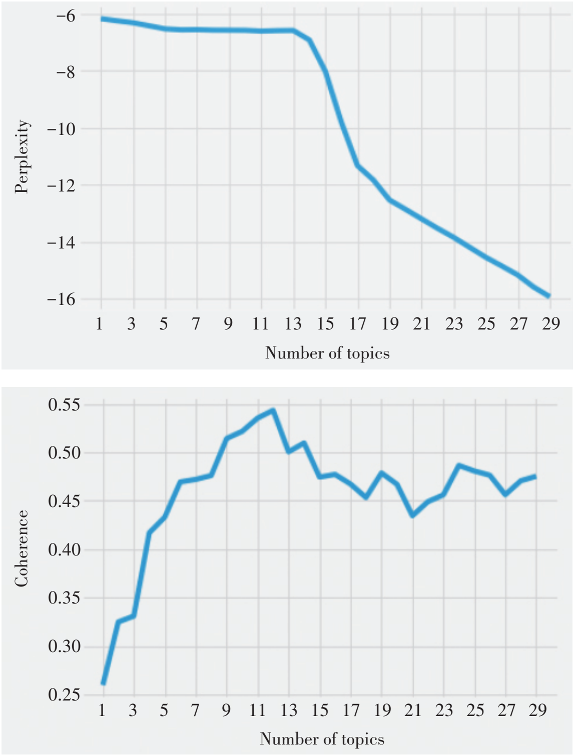


SSE: sum of square error

▲**Figure 2. Principal component analysis (PCA) results, K-means SSE, and K-means silhouette curves using bag of words (BoW) model**

the topic, symptom, and the combined token list. To determine the performance of the optimal number of topics, we compare different perplexity scores and coherence scores when applying different topic numbers. We select the topic number at the inflection point where the perplexity curve or the coherence curve turns.

Fig. 3 shows the results of the cloud infrastructure ticket data using LDA with 1 – 30 topics. Based on the characteristics of the curves, we can select the number of topics to be 14. The top ten keywords and the probabilities for each of the topics are shown in Table 2.



LDA: latent dirichlet allocation

▲**Figure 3. LDA perplexity and coherence curves**

▼**Table 2. Topics and keywords in each topic**

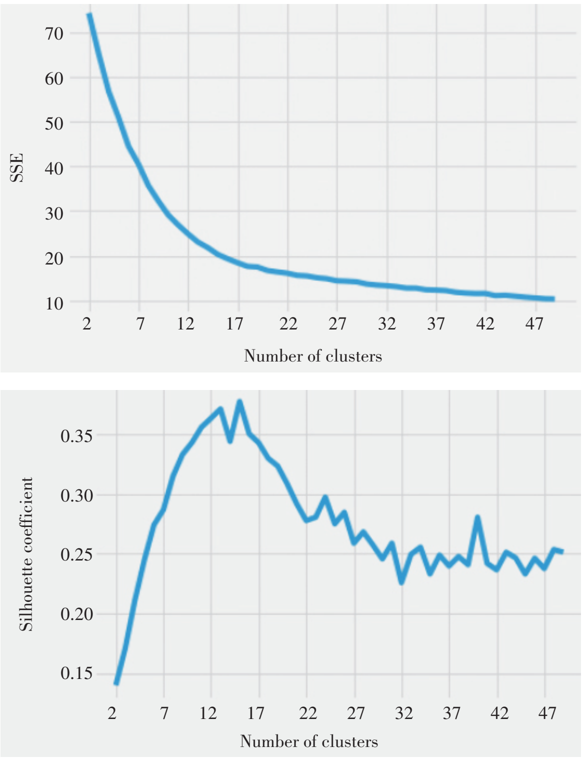
Topic	Keywords and Probabilities in Each Topic									
Topic 0	interface	daisy	NIC	file	maintain	virtualization	platform	zone	mode	increase
	0.107	0.055	0.034	0.033	0.025	0.023	0.023	0.022	0.021	0.020
Topic 1	OS	start	capacity	install	stack	service	blade	VM	mac	down
	0.113	0.090	0.072	0.054	0.029	0.027	0.025	0.017	0.017	0.017
Topic 2	provider	performance	device	login	director	memory	query	recover	bandwidth	occupy
	0.218	0.048	0.038	0.037	0.037	0.031	0.029	0.026	0.021	0.019
Topic 3	dvs	gateway	project	support	manage	business	resource	pool	add	thread
	0.081	0.041	0.038	0.038	0.033	0.032	0.031	0.028	0.027	0.023
Topic 4	network	blocked	power on	udm	update	management	information	change	packet loss	endpoint
	0.148	0.061	0.035	0.032	0.03	0.029	0.024	0.023	0.022	0.013

NIC: network interface card VM: virtual machine

We evaluate the probability of topic appearance for each incident ticket and then cluster the topics using *K*-means. Fig. 4 shows the SSE curve and silhouette coefficient curve respectively. The curves demonstrate more significant turning points than the ones using the BoW model, indicating the LDA model is more suitable for incident ticket clustering.

With 14 topics, the SSE of LDA-allocated tokens is about 20, while the SSE of BoW is about 215, an order of higher magnitude. Though it is unpragmatic to map the SSE score to the exact accuracy, the lower the score the more accurate the prediction. Similarly, the silhouette coefficient for LDA results with 14 topics are about 0.37, compared with less than 0.10 using BoW models. As the score measures how apart of the cluster ranging from −1 to 1, a value close to 1 indicates clearly distinguished clusters.

Table 3 shows the titles of tickets in one cluster. Storage related problems consist of a majority of the tickets, especially during upgrade and backup stages. The next is networking issues.



SSE: sum of square error

▲**Figure 4. K-means SSE and silhouette curves using latent Dirichlet allocation (LDA) model**

4.4 Incident Ticket Classification and Prediction

Our ticket clustering experiments reveal that incident tickets do have clustering characteristics. In order to take full advantage of prior knowledge, e.g., to assign coming tickets to the same department which has resolved similar ones before, we study the classification and prediction of incident tickets in this section. We use a similar dataset with more fields, including ticket ID, ticket description, resolution, resolution groups, categories, sub-categories, and components. After removing null values, the categories and record numbers are shown in Table 4.

▼**Table 3. Samples of title descriptions in one cluster**

ID	Title Description	ClusterID
BC20200229-0052	Backup of version 6.10.10.08 failed during upgrading	5
BC20200307-0094	Problems of 3.17.15.06P2 trusted resource pool	5
BC20200309-0105	Backup of 6.10.10.P8tecs6.0 environment failed	5
BC20200402-0034	Nova service failed due to version 3.17.15.06 license problem during vHSS capacity increase	5
BC20200404-0051	Provider MariaDB failed to start after rebooting one of the control node in group one	5
BC20200409-0023	Keystone service abnormal in the 5GC test node in 3.17.14 control environment	5
BC20200411-0045	Two of the disks failed when batch creating 32 35T cloud disks reporting not sufficient space	5
BC20200507-0012	Mutual trust failed during Northeast 3.17.15.06P4 upgrading	5
BC20200511-0058	Part of the related information not updated after changing configuration of 3.17.15.06P3 on Daisy	5
BC20200520-0066	One of the VMs failed during start reporting volume not found after NFVINMA1Z station upgrade	5
BC20200603-0025	Failure of V03.17.15.06P4HP3 upgrading	5
BC20200603-0027	Multi-node upgrade failed due to 3.17.15.06P4 17.15.06P4HP3 mutual trust lost	5
BC20200618-0066	3.17.15.07_T2-daisy upgrade failed	5
BC20200629-0247	V3.17.14.P2Provider auto-backup service hang	5
BC20200630-0257	Network issues from two VMs on 3.17.15.06P4HP3	5
BC20200702-0060	Mirror file upload failed after V03.17.15.07T2-Provider upgrade	5
BC20200705-0002	Nova service down after 3.17.15.06P4HP3 upgrade	5
BC20200710-0121	3.17.15.08-OS distribution failed reporting mutual-trust issue	5
BC20200710-0123	40 VMs in shutoff status using 3.17.15.08 startup script	5
BC20200716-0095	Not able to apply new license after 3.17.15.06P6-license failed	5
BC20200722-0207	Tenant resource abnormal after tenants with the same name created	5

vHSS: virtual home subscriber server VM: virtual machine

There are 115 sub-categories and 49 of them contain 1 000 records or more. The 49 sub-categories consist of 96% of the total tickets, and 30 of them contain 3 000 records or more consisting of 87% of the total records. When it comes to components, there are 663 in total, among which there are 88 items with more than 1 000 records accounting for 79% of the total amount, and 34 items with more than 3 000 records ac-

▼Table 4. Categories and record numbers

Category	Number of Records
Infrastructure	177 040
Operation product line	64 869
OA product line	55 570
EPMS intelligent service	22 454
iCenter application	19 849
PLM product line	15 870
AIOps group	14 121
Technical platform	1 232
Others	171
IT Wizard	19
Operation NOC	17
Network	5
Communication	2
Middleware	1
Security	1

AIOps: artificial intelligence for IT operations
EPMS: enterprise performance management system
NOC: network operations center
OA: office automation
PLM: product lifecycle management

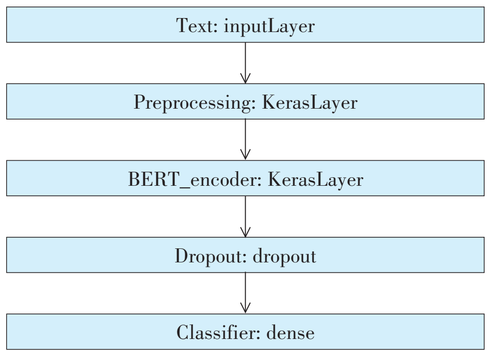
counting for 54% of the total amount.

In order to achieve fine granularity of the classification, we use the combination of sub-categories and components as the label. There are 29 top labels with more than 3 000 records.

We compare multiple classification algorithms including TF-IDF, LDA and BERT. As expected, BERT achieved the best precision and recall for the same dataset. Both TF-IDF and LDA with the regression model yield a prediction accuracy of less than 80%. We build the incident classification model based on BERT which is shown in Fig. 5.

1) Architecture of our model

- The input layer is a text layer with preprocessed incident description text.
- The preprocessing layer is a Chinese processing model devised by Google (suited for the BERT model). Every ticket text is transformed into 3 vectors: input_word_ids, input_mask and input_type_ids with 128 dimensions respectively. Input_word_ids denotes the ID of the word. The lost elements of input_word_ids vector are filled with 0. For the corresponding numbers in an input_mask vector, they should be 1 while the remaining elements are 0. Input_type_ids can clarify different sentences. In this classification study, we set all of its elements to 0.



BERT: bidirectional encoder representation from transformers

▲Figure 5. BERT classification network architecture

- BERT_encoder is an advanced BERT model devised by Google. BERT_encoder has 12 layers (bert_zh_L-12_H-768_A-12l) and the output of the BERT_encoder consists of pooled_output (each text corresponds to a vector of 768 elements), sequence_output (each word in each text corresponds to a vector of 768 elements) and encoder_outputs (output of inner units). We only focus on pooled_output in this experiment.
- The dropout layer aims at avoiding overfitting. The probability of dropout is set to 0.1.
- The classifier layer is a fully connected layer that outputs the probability of each ticket belonging to a certain classification in the labels.

2) Training and testing data preparation

We use the following steps as data preprocessing to generate training and testing data:

- Delete all the incident tickets containing null value category information or empty ticket descriptions.
- Modify the classification labels into lowercase and delete the redundant blank space. This operation is devised from observing the original data, where some categories and items are generally the same but only differ in lowercase and uppercase. For example, iCenter and Icenter.
- Delete tickets with ambiguous items and category labels like “other, others, to be classified, and other pending problems”.
- Merge the item and category labels in the form of component. category such as intelligent maintenance.itsp serve website.
- After the merging operation, delete labels and their incident tickets data whose statistic number is less than the threshold (we set 3 000 in this experiment).
- Remove HTML formatting and redundant space (including line feed punctuation) from the incident description texts. For the English content, all the letters are also put in lowercase.
- Shuffle the resulting incident data. 70% of the dataset is utilized as the training set and the remaining 30% is used as the test set.
- Each classification label and its quantity of relevant incident tickets are given in Table 5 (29 classification labels with more than 3 000 records respectively are reserved).

As a result, 103 094 incident tickets are identified as training data and 44 183 incident tickets are collected as test data.

For training the model, we adopt the Sparse Categorical Crossentropy as the loss function, Sparse Categorical Accuracy for accuracy measurement and optimize the model with AdamW. The experiment sets the initial learning rate to 3e−5 and the epoch to 5. The original training data are partitioned

▼**Table 5. Top labels (combination of sub-categories and components) and record numbers**

Classification Label	Numbers of Records	ID
AIOps - itsp service website	16 724	0
desktop cloud - linux desktop cloud	11 222	1
desktop cloud - OS issues	7 162	2
PC side zmail - operation issues	7 110	3
ifol finance - oerp enterprise resource planning	6 543	4
desktop cloud - intranet client-side login	6 531	5
ifol finance - fol finance online	6 381	6
network proxy - usage issues	5 511	7
iscp supply chain - iwms.wms cloud storage	5 086	8
iccp customer relationship - msm marketing	4 994	9
iccp customer relationship - ccg contract configuration	4 779	10
PC side zmail - account issues	4 622	11
iscp supply chain - iwms.mcs manufacture management	4 207	12
im instant message - usage issues	4 156	13
PC side zmail - account creation & login	4 043	14
ifol finance - cms contract management web	4 019	15
uds failure - security check	3 930	16
ifol finance - cms contract management form	3 886	17
icenter - ts team coordination	3 878	18
desktop cloud - blue or black screen	3 872	19
engineering domain - cca cloud code review	3 584	20
desktop cloud - client side login failed	3 551	21
OS issues - installation	3 531	22
ibcp human resource - hol online	3 258	23
iscp supply chain - isrm.srm supplier management	3 254	24
Mobil application - icenter Mobile side	3 234	25
individual network issue - restriction	3 213	26
ibcp human resource - elearning academy	3 178	27
uds failure - usage issues	3 008	28

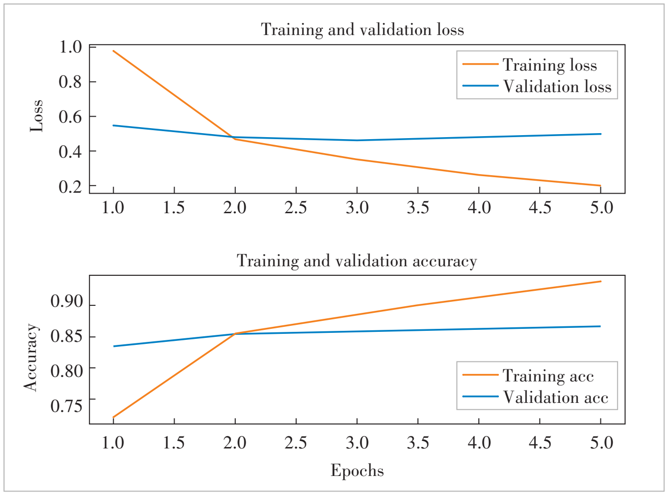
into a training set and a validation set at the ratio of 9∶1 in this pre-training procedure (i.e., the number of incident tickets used for model training is the number of preprocessed incident tickets × 70% × 90%).

Fig. 6 shows the training loss, training accuracy, validation loss, and validation accuracy of each epoch.

To verify our model after pretraining, we perform classification tasks on the test set. The assessment results are illustrated in Table 6. The overall precision is up to 86%. The confusion matrix of prediction results is shown in Table 7. The number in cell (*i*, *j*) denotes the number of tickets, the labels of which are *i* but predicted to be *j* in this model. Therefore, the numbers of correctly classified incident tickets lie on the diagonal while the number lying off the diagonal shows the discrepancies in classification.

$$\frac{N(i,j)}{\sum_{k=0}^{28} N(i,k)} \geqslant 5\%, i \neq j . \tag{1}$$

From Table 8, we can see that a majority of classification error cases occurs among different items of the same sub-categories. For example, it is confusing for the model to classify the items of some sub-categories like desktop cloud, PC side zmail incidents, ifol finance and uds failure. In addition, 7.4% tickets of OS issues - installation are wrongly predicted



▲**Figure 6. Training loss and accuracy vs validation loss and accuracy**

▼**Table 6. Prediction accuracy on test data**

	Precision	Recall	F1-score	Support
0	0.91	0.91	0.91	4 951
1	0.85	0.85	0.85	3 289
2	0.67	0.65	0.66	2 071
3	0.80	0.85	0.83	2 146
4	0.88	0.88	0.88	1 931
5	0.73	0.75	0.74	1 974
6	0.93	0.93	0.93	1 872
7	0.91	0.93	0.92	1 619
8	0.93	0.92	0.93	1 543
9	0.91	0.93	0.92	1 470
10	0.91	0.90	0.91	1 500
11	0.84	0.80	0.82	1 381
12	0.90	0.91	0.91	1 231
13	0.91	0.93	0.92	1 266
14	0.80	0.82	0.81	1 188
15	0.80	0.80	0.80	1 206
16	0.89	0.90	0.90	1 228
17	0.79	0.77	0.78	1 159
18	0.93	0.94	0.93	1 119
19	0.67	0.71	0.69	1 173
20	0.95	0.95	0.95	980
21	0.87	0.89	0.88	1 046
22	0.89	0.88	0.89	1 060
23	0.90	0.89	0.90	925
24	0.91	0.91	0.91	978
25	0.95	0.93	0.94	994
26	0.87	0.77	0.81	960
27	0.95	0.96	0.95	955
28	0.79	0.69	0.74	968
Accuracy			0.86	44 183
Macro avg	0.86	0.86	0.86	44 183
Weighted avg	0.86	0.86	0.86	44 183

▼Table 7. Confusion matrix of prediction results

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
0	4 520	5	6	22	4	14	8	20	1	1	7	41	2	28	42	5	16	1	26	2	0	5	58	24	7	35	13	4	34
1	12	2 789	166	24	1	118	1	16	1	3	0	0	1	26	16	1	0	0	4	60	4	13	5	1	4	0	7	4	12
2	8	216	1 352	6	0	233	0	2	0	2	1	0	1	5	2	1	2	1	0	203	1	15	8	0	0	0	7	0	5
3	17	21	6	1 826	1	1	3	5	0	1	2	139	1	1	106	1	0	1	6	1	1	0	1	1	0	0	0	3	1
4	2	2	0	0	1 706	0	49	0	19	2	2	1	56	1	1	37	0	33	0	0	7	0	0	4	6	0	1	1	1
5	10	70	242	2	0	1 473	0	0	0	0	1	0	0	1	2	0	3	0	1	113	2	47	0	0	0	0	6	0	1
6	6	5	0	3	61	0	1 745	1	4	5	3	1	2	0	0	9	0	2	3	0	2	0	0	6	9	0	0	3	2
7	21	27	1	3	0	0	2	1 510	2	0	2	0	0	1	4	1	1	0	5	0	3	1	1	1	2	0	10	5	16
8	4	1	1	2	22	0	8	2	1 421	1	6	1	8	0	0	15	1	11	4	0	0	0	0	0	27	1	1	2	1
9	7	0	1	2	2	0	6	1	3	1 374	39	1	0	1	1	7	0	5	5	0	1	0	2	7	3	0	0	2	0
10	5	2	1	3	3	0	2	3	2	66	1 352	0	5	3	1	7	2	30	1	0	7	0	1	0	2	0	0	1	1
11	28	6	1	196	1	1	1	0	1	0	1	1 103	1	1	27	0	1	1	2	0	1	0	0	1	2	1	1	2	1
12	6	3	2	0	29	0	2	0	25	3	4	0	1 117	1	0	4	0	12	1	0	4	0	3	4	8	0	2	0	1
13	14	24	3	7	1	0	3	1	0	1	0	3	2	1 173	2	4	0	1	1	0	1	1	4	5	4	2	2	2	5
14	31	16	2	125	0	5	1	2	0	0	1	12	0	3	969	0	0	0	2	0	1	2	0	3	1	2	4	0	6
15	1	0	0	3	51	0	7	1	10	13	7	0	1	3	0	966	0	132	6	0	1	0	0	0	3	0	0	1	0
16	30	4	6	2	1	4	0	1	0	0	0	2	2	3	0	0	1 107	1	0	0	3	1	4	1	1	0	4	1	50
17	0	0	0	0	40	0	0	0	15	13	29	1	23	1	1	140	0	892	0	0	2	0	0	0	1	0	1	0	0
18	19	0	1	12	2	0	6	1	2	1	0	4	1	3	1	0	0	2	1 050	0	1	3	0	4	2	2	0	1	1
19	1	42	174	4	0	93	0	0	0	1	0	0	1	2	0	0	0	0	0	832	1	17	3	0	0	0	2	0	0
20	5	9	1	6	3	0	2	3	2	2	4	1	2	3	0	0	2	1	1	0	928	0	0	2	1	0	2	0	0
21	7	11	15	1	0	50	0	3	1	0	0	0	0	1	2	0	0	0	0	17	0	926	2	1	0	1	4	0	4
22	78	5	9	2	2	1	0	0	0	1	2	0	3	2	1	0	6	1	0	6	0	1	933	0	1	0	2	1	4
23	42	2	0	2	2	1	12	0	0	3	2	3	0	6	0	0	0	0	7	0	0	1	0	826	5	5	2	2	3
24	6	4	0	4	6	1	12	1	16	5	9	1	7	2	2	2	0	5	4	0	0	0	2	1	886	1	0	1	0
25	35	5	0	3	0	0	0	0	0	0	1	1	0	3	1	0	0	0	4	0	0	0	1	7	1	927	0	4	1
26	25	9	18	3	1	28	1	71	0	2	0	1	0	1	6	1	2	1	1	6	5	5	6	2	0	0	736	3	26
27	4	4	3	4	0	1	5	2	0	1	2	2	0	1	1	0	0	0	0	0	0	0	1	6	1	0	3	914	0
28	44	11	5	7	1	5	6	6	1	1	1	1	1	12	17	0	99	0	1	2	1	22	8	6	0	1	40	3	666

▼Table 8. Sample labels of classification error

Ground Truth Category/Erroneous Category	Total Samples	Error Samples	
desktop cloud - linux desktop cloud	3 289		
• desktop cloud - OS issues		166	5.0%
desktop cloud - OS issues	2 071		
• desktop cloud - linux desktop cloud		216	10.4%
• desktop cloud - intranet client-side login		233	11.3%
• desktop cloud - blue or black screen		203	9.8%
PC side zmail - operation issues	2 146		
• PC side zmail - account issues		139	6.5%
desktop cloud - intranet client-side login	1 974		
• desktop cloud - OS issues		242	12.3%
• desktop cloud - blue or black screen		113	5.7%
PC side zmail - account issues	1 381		
• PC side zmail - account creation & login		196	14.2%
PC side zmail - account creation & login	1 188		
• PC side zmail - operation issues		125	10.5%
ifol finance - cms contract management web	1 206		
• ifol finance - cms contract management form		132	10.9%
ifol finance - cms contract management form	1 159		
• ifol finance - cms contract management web		140	12.1%
desktop cloud - blue or black screen	1 173		
• desktop cloud - OS issues		174	14.8%
• desktop cloud - intranet client-side login		93	7.9%
OS issues-installation	1 060		
• AIOps - itsp service website		78	7.4%
individual network issue - restriction	960		
• network proxy - usage issues		71	7.4%
uds failure-usage issues	968		
• uds failure - sercurity check		99	10.2%

AIOps: artificial intelligence for IT operations

to be AIOps-itsp service website and 7.4% tickets of individual network issues – restriction are wrongly predicted to be network proxy-usage issues.

5 Conclusions

In this paper, we demonstrate the semantic characteristics of problem and incident tickets. Taking the ticket data from a real production Cloud environment, we compare different text mining techniques. LDA and K-Means are applied to show the ticket clusters. We use BERT as the deep learning framework with fine-tuning to build a resolution department matching system. Using sub-category and component fields in the ticket description, our classification model achieves 86% accuracy when predicting the best match department to resolve the ticket.

Reference

[1] FORELL T, MILOJICIC D, TALWAR V. Cloud management: challenges and opportunities [C]//IEEE International Symposium on Parallel and Distributed Processing Workshops and PhD Forum. IEEE, 2011: 881 – 889. DOI: 10.1109/IPDPS.2011.233

[2] MARTIN-FLATIN J P. Challenges in cloud management [J]. IEEE cloud computing, 2014, 1(1): 66 – 70. DOI: 10.1109/MCC.2014.4

[3] PEREIRA R, DA SILVA M M. Towards an integrated IT governance and IT management framework [C]//The 16th International Enterprise Distributed Object Computing Conference. IEEE, 2012: 191 – 200. DOI: 10.1109/EDOC.2012.30

[4] FARIA N E, SILVA M M. Selecting a software tool for ITIL using a multiple criteria decision analysis approach [EB/OL]. [2022-12-10]. https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1241&context=isd2014

[5] BOTEZATU M M, BOGOJESKA J, GIURGIU I, et al. Multi-view incident ticket clustering for optimal ticket dispatching [C]//The 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015: 1711 – 1720. DOI: 10.1145/2783258.2788607

[6] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [EB/OL]. [2022-12-10]. https://arxiv.org/abs/1810.04805.pdf

[7] ROELLEKE T, WANG J. TF-IDF uncovered: a study of theories and probabilities [C]//The 31st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2008: 435 – 442. DOI: 10.1145/1390334.1390409

[8] JELODAR H, WANG Y L, YUAN C, et al. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey [J]. Multimedia tools and applications, 2019, 78(11): 15169 – 15211. DOI: 10.1007/s11042-018-6894-4

[9] MO Z L. Stopwords [EB/OL]. (2019-12-18) [2022-12-10]. https://github. com/goto456/stopwords/blob/master/baidu_stopwords.txt

[10] LARS Y C. Stopwords [EB/OL]. (2011-12-06) [2022-12-10]. https://gist.github.com/larsyencken/1440509

[11] DU M, LI F F, ZHENG G N, et al. DeepLog: anomaly detection and diagnosis from system logs through deep learning [C]//The 2017 ACM SIGSAC Conference on Computer and Communications Security. ACM, 2017: 1285 – 1298. DOI: 10.1145/3133956.3134015

[12] ZHANG X, XU Y, LIN Q W, et al. Robust log-based anomaly detection on unstable log data [C]//Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering. ACM, 2019: 807 – 817. DOI: 10.1145/3338906.3338931

[13] LIN J Y, ZHANG Q, BANNAZADEH H, et al. Automated anomaly detection and root cause analysis in virtualized cloud infrastructures [C]//IEEE/IFIP Network Operations and Management Symposium. IEEE, 2016: 550 – 556. DOI: 10.1109/NOMS.2016.7502857

[14] HARUTYUNYAN, A N, GRIGORYAN N M, POGHOSYAN A V, et al. Intelligent troubleshooting in data centers with mining evidence of performance problems [EB/OL]. (2020-09-20) [2022-12-10]. https://www.researchgate.net/publication/344251115_Intelligent_Troubleshooting_in_Data_Centers_with_Mining_Evidence_of_Performance_Problems

[15] BOU NASSIF A, ABU TALIB M, NASIR Q, et al. Machine learning for cloud security: a systematic review [J]. IEEE access, 2021, 9: 20717 – 20735. DOI: 10.1109/ACCESS.2021.3054129

[16] GRZONKA D, JAKÓBIK A, KOŁODZIEJ J, et al. Using a multi-agent system and artificial intelligence for monitoring and improving the cloud performance and security [J]. Future generation computer systems, 2018, 86: 1106 – 1117. DOI: 10.1016/j.future.2017.05.046

[17] SHAO Q H, CHEN Y, TAO S, et al. Efficient ticket routing by resolution sequence mining [C]//Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 2008: 605 – 613. DOI: 10.1145/1401890.1401964

[18] AGARWAL S, SINDHGATTA R, SENGUPTA B. SmartDispatch: enabling efficient ticket dispatch in an IT service environment [C]//The 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012: 1393 – 1401. DOI: 10.1145/2339530.2339744

[19] LIN D, RAGHU R, RAMAMURTHY V, et al. Unveiling clusters of events for alert and incident management in large-scale enterprise it [C]//The 20th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 2014: 1630 – 1639. DOI: 10.1145/2623330.2623360

[20] MANI S, SANKARANARAYANAN K, SINHA V S, et al. Panning requirement nuggets in stream of software maintenance tickets [C]//The 22nd ACM SIGSOFT International Symposium on Foundations of Software Engineering. ACM, 2014: 678 – 688. DOI: 10.1145/2635868.2635897

[21] AGARWAL S, AGGARWAL V, AKULA A R, et al. Automatic problem extraction and analysis from unstructured text in IT tickets [J]. IBM journal of research and development, 2017, 61(1): 41 – 52. DOI: 10.1147/JRD.2016.2629318

[22] JAN E E, CHEN K Y, IDÉ T. Probabilistic text analytics framework for information technology service desk tickets [C]//IFIP/IEEE International Symposium on Integrated Network Management (IM). IEEE, 2015: 870 – 873. DOI: 10.1109/INM.2015.7140397

Biographies

FENG Hailin received his PhD in computer science from the University of Science and Technology of China in June 2007. Since 2007, he has been working in the School of Information Engineering of Zhejiang A&F University, China. From 2013 to 2014, he was a visiting professor at Forest Products Laboratory, USA. He is currently a professor in the School of Mathematics and Computer Science and School of Information Engineering of Zhejiang A&F University. His main interest areas include computer vision, intelligent information processing, and Internet of Things.

HAN Jing (han.jing28@zte.com.cn) received her master's degree from Nanjing University of Aeronautics and Astronautics, China. She has been with ZTE Corporation since 2000, where she worked on 3G/4G key technologies from 2000 to 2016 and has become a technical director responsible for intelligent operation of cloud platforms and wireless networks since 2016. Her research interests include machine learning, data mining, and signal processing.

HUANG Leijun received his PhD in computer science from George Mason University, USA in 2008. Since 2010, he has been working in the School of Information Engineering of Zhejiang A&F University, China. He is currently a lecturer in the School of Mathematics and Computer Science. His main interest areas include computer networks, Internet of Things and data mining.

SHENG Ziwei received her BS degree in software engineering from Huazhong University of Science and Technology, China in 2022. She is currently pursuing her MS degree in electrical and computer engineering at Carnegie Mellon University, USA. In her master's program, she primarily focuses on the fields of engineering development and system design. Her ultimate goal is to advance technology and foster innovation in these domains.

GONG Zican received his master's degree in professional computing and artificial intelligence from the Australian National University in 2019. He has been a machine learning engineer in ZTE Corporation since 2020. His research interests include machine learning, professional computing and system architecture.

A Hybrid Five-Level Single-Phase Rectifier with Low Common-Mode Voltage

TIAN Ruihan[1], WU Xuezhi[1], XU Wenzheng[1],
ZUO Zhiling[2], CHEN Changqing[2]

(1. Beijing Jiaotong University, Beijing 100044, China;
2. ZTE Corporation, Shenzhen 518057, China)

DOI: 10.12142/ZTECOM.202304010

https://kns.cnki.net/kcms/detail/34.1294.TN.20231114.1558.002.html, published online November 16, 2023

Manuscript received: 2023-03-19

Abstract: Rectifiers with high efficiency and high power density are crucial to the stable and efficient power supply of 5G communication base stations, which deserves in-depth investigation. In general, there are two key problems to be addressed: supporting both alternating current (AC) and direct current (DC) input, and minimizing the common-mode voltage as well as leakage current for safety reasons. In this paper, a hybrid five-level single-phase rectifier is proposed. A five-level topology is adopted in the upper arm, and a half-bridge diode topology is adopted in the lower arm. A dual closed-loop control strategy and a flying capacitor voltage regulation method are designed accordingly so that the compatibility of both AC and DC input is realized with low common voltage and small passive devices. Simulation and experimental results demonstrate the effectiveness and performance of the proposed rectifier.

Keywords: multilevel rectifier; 5L-ANPC; low common-mode voltage; AC-DC hybrid input

Citation (Format 1): TIAN R H, WU X Z, XU W Z, et al. A hybrid five-level single-phase rectifier with low common-mode voltage [J]. ZTE Communications, 2023, 21(4): 78–84. DOI: 10.12142/ZTECOM.202304010

Citation (Format 2): R. H. Tian, X. Z. Wu, W. Z. Xu, et al., "A hybrid five-level single-phase rectifier with low common-mode voltage," *ZTE Communications*, vol. 21, no. 4, pp. 78–84, Dec. 2023. doi: 10.12142/ZTECOM.202304010.

1 Introduction

With the development of 5G networks, reliable power supply is the key to ensuring the safe and stable operation of communication systems. Higher power efficiency, lower power noise, and higher stability and reliability are required. Therefore, the application of power factor correction (PFC) rectifiers in communication power supply has attracted wide attention[1–3], mainly to meet the power quality requirements of 5G communication base stations as well as the need for alternating current (AC) and direct current (DC) input switching[4].

At present, many researchers study multilevel converters[5–6], which can control the output terminals of different DC power supplies connected in series through a specific circuit topology. With the change of different switching states of the circuit, the multi-step wave can be equivalent to a sine wave. Compared with the traditional two-level converter, multilevel converters have the following advantages: 1) The voltage stress of semiconductor switching devices is reduced to achieve a higher level of voltage and power output; 2) it has high output power quality, smaller dv/dt, low total harmonic distortion (THD) and electromagnetic interference; 3) it can operate at both the fundamental wave and high frequency switching frequency, reducing the switching loss of power switches and improving system efficiency. Thanks to the above advantages, multilevel rectifiers are widely used in power supply systems, power factor correction, battery energy storage systems and other applications[7–8]. However, the multilevel converter still has some shortcomings, one of which is that the number of switching devices increases exponentially with the increase of the level numbers, and each switch requires a gate driving circuit, which complicates the control strategy of the system and increases the cost[9]. In order to pursue better performance, researchers continue to innovate and improve the multilevel topology.

Among various multilevel converters, the five-level active neutral-point-clamped (5L-ANPC) topology is a topology with high practicability and good economy. Compared with the cascaded H-bridge (CHB) converter, only one DC source is needed. Compared with the neutral-point-clamped (NPC) converter, the number of clamping devices is reduced. Nevertheless, the number of redundant switches is increased by introducing the flying capacitor. The DC link is divided into two parts, which reduces the difficulty of voltage equalization of the DC capacitor and realizes the capacitor voltage balance through a certain control algorithm. Compared with the flying-capacitor (FC) converter, 5L-ANPC greatly reduces the amount of capacitance used and has great advantages[10]. In order to suppress the common-mode voltage better, the combination of upper and lower bridge arms is adopted to further increase the number of levels and reduce d*v*/d*t* in this paper. Otherwise, if the 5L-ANPC topology is adopted for both upper and lower bridge arms, 16 switches are required. The number of switches is too large to meet the requirements of high power density. This paper presents a hybrid rectifier topology by combining the upper bridge arm 5L-ANPC and lower bridge arm diodes.

The rest of the paper is organized as follows. In Section 2, the operating characteristics and the control strategy of the 5L-ANPC hybrid rectifier are introduced, including FC voltage regulation. In Section 3, simulations and experiments verify the effectiveness and performance of the strategy proposed in Section 2. Finally, the conclusion is given in Section 4.

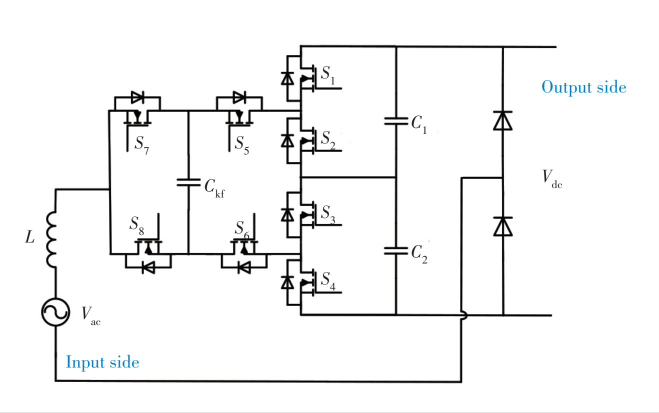
2 Working Principle

2.1 Topology Analysis

A traditional multilevel topology is difficult to apply in five or higher levels because of the inability to control the point voltage in the bus bar and the excessive number of clamping elements. The 5L-ANPC has aroused wide concern after it was proposed. It can improve the output voltage level, using fewer devices when outputting the same number of levels, and its capacitor voltage equalizing control algorithm is relatively simple, which has great application potential in switching power supply.

The AC/DC rectifier circuit needs to realize power factor correction and suppress common-mode current. More levels can be achieved through the combination of upper and lower bridge arms, but it conflicts with the demand for high power density. Therefore, the topological structure of five levels on the upper arm and the diode on the lower arm is adopted, as shown in Fig. 1.

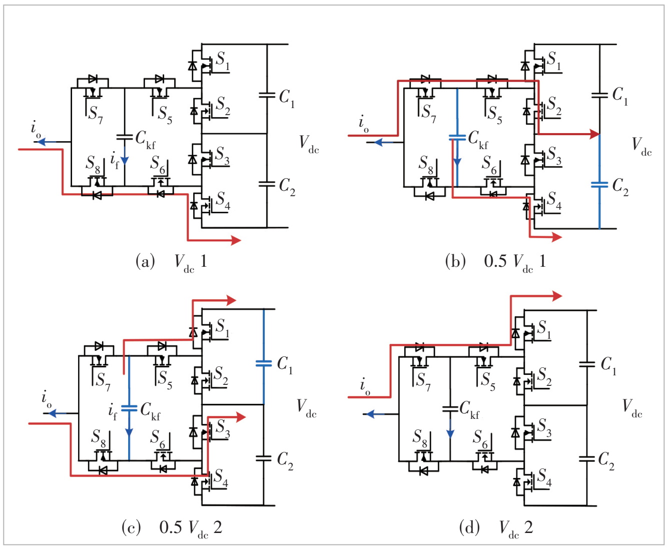
The upper bridge arm is a single-phase 5L ANPC converter topology, consisting of 8 switches with inverse shunt diodes, an FC C_f, and two supporting capacitors C_1 and C_2 on the DC side. The first part consists of two supporting capacitors and four switching devices, which are used to clamp the busbar voltage. The latter part uses the FC structure for multilevel output. Assuming that the voltage at the DC side is 4*E*, the conditions for the topology to work properly are as follows. 1) The voltage of the two support capacitors at the DC side is basically the same, that is, $U_{C_1} = U_{C_2} = U_{dc}/2 = 2E$. 2) The voltage of the flying span capacitor is ensured to be basically stable, i.e., $U_{Cf} = U_{dc}/4 = E$.



▲Figure 1. Topology of hybrid 5L single-phase rectifier (adopting two-stage bridge scheme)

The corresponding switching devices of 5L ANPC converter topology are complementary, and their actions must follow the following switching principles[8]: 1) The switching devices S_1 and S_3, S_2 and S_4 must be turned on or off at the same time, respectively. The two groups are complementary and operate at power frequency; 2) The switching devices S_5 and S_7, S_6 and S_8 should be complementary respectively and operate at the switching frequency; 3) When switching mode of switches is selected, two pairs of switch devices operating at the same time is usually avoided.

The positive direction of the output current i_o of the converter bridge arm and the FC current i_f are denoted in Fig. 2. All switching states and their related currents can be obtained in Table 1. It can be seen that there are eight different switching states from V_1 to V_8, and the rectifier can produce five volt-



(a) V_{dc} 1

(b) 0.5 V_{dc} 1

(c) 0.5 V_{dc} 2

(d) V_{dc} 2

▲Figure 2. Status of the direct current (DC) input switching

▼Table 1. Five-level active neutral-point-clamped (5-L ANPC) converter switching

	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	V_0	i_o	i_f
V_1	0	0	1	1	0	1	0	1	$-2E$	0	0
V_2	0	1	1	0	0	1	0	1	$-E$	i_o	0
V_3	1	0	0	1	0	1	0	1	$-E$	$-i_o$	i_o
V_4	1	1	0	0	0	1	0	1	0	0	i_o
V_5	0	0	1	1	1	0	1	0	0	0	i_o
V_6	0	1	1	0	1	0	1	0	E	i_o	i_o
V_7	1	0	0	1	1	0	1	0	E	$-i_o$	0
V_8	1	1	0	0	1	0	1	0	$2E$	0	0

age levels under the equilibrium condition, namely 0.5 V_{dc}, 0.25 V_{dc}, 0, −0.25 V_{dc} and −0.5 V_{dc}. When the output level is $-E$, 0 and E, there are two redundant switching states respectively, and when the output level is $-E$ and E, the corresponding two switching states, namely (V_2, V_3) and (V_6, V_7), have opposite effects on the charging and discharging of the FC voltage. The voltage equalizing control of FC can be realized by adjusting the two switching states properly.

For a single-phase rectifier, common-mode voltage is defined as $V_{\mathrm{cm}}=(V_a+V_b)/2$, where V_a, and V_b are the voltages of bridge arms A and B, and the common-mode current has a direct relationship with the amplitude of common-mode voltage and the step slope $\mathrm{d}V_{\mathrm{cm}}/\mathrm{d}t$. Therefore, effectively reducing the amplitude and jump slope of common-mode voltage can effectively reduce the common-mode current.

The proposed five-level topology has a total of eight switching devices, which need to provide eight +15 V/−5 V on/off signals to ensure the safety and accuracy of the driving signals. The vertical switches S_1 – S_4 are operated at the power frequency and can be isolated from the main circuit by an isolation transformer. The switches S_5 – S_8 are operated at the high frequency, where the upper and lower pairs are conducted complementarily. To simplify the drive circuit and reduce the cost and complexity of hardware implementation, the bootstrap drive circuit can be adopted to split the power supply into two groups to reduce the number of independent isolated power supplies.

2.2 DC Input Case

The proposed converter also supports DC input voltage. For the DC input case, the operation of this converter is downgraded to a three-level boost circuit. In order to preserve the five-level scheme under AC input and avoid the problem of large common mode inductance required by the two-level scheme for DC, the three-level control strategy for DC is adopted after research.

In the proposed three-level control strategy, V_{Ckf} shall be equal to (1/2)V_{dc}, namely 200 V. Voltage levels 0, 0.5 V_{dc}, and V_{dc} are used to synthesize voltage. When duty cycle $d<0.5$, 0 and 0.5 V_{dc} are used to synthesize voltage; when $0.5<d<1$, 0.5 V_{dc} and V_{dc} are used to synthesize voltage. There are four switching states as shown in Fig. 2, among which 0.5 V_{dc} corresponds to two switching states. The FC is used to realize time-sharing in parallel with the upper and lower capacitors, so as to achieve voltage balancing between capacitors C_1 and C_2. The on-off time of the two switching states is the same and the carrier phase-shift modulation strategy, phase-shift pulse width modulation (PS-PWM), is adopted. For the DC input case, the three-level switching states are shown in Table 2.

▼Table 2. 3-L boost switching table in a direct current (DC) case

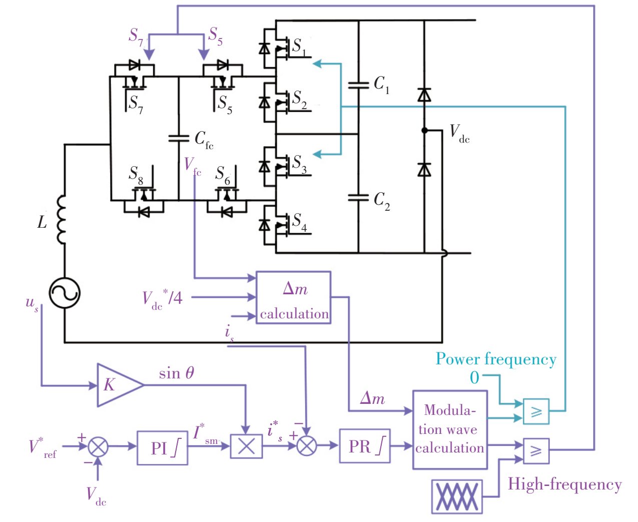
	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8
V_1	0	1	0	1	0	1	0	1
V_2	0	1	0	1	1	1	1	0
V_3	1	0	1	0	1	1	0	1
V_4	1	0	1	0	1	0	1	0

2.3 Control Strategy

Regarding the design of a single-phase five-level rectifier with excellent performance, it is necessary to meet the following aspects: achieving unit power factor operation and ensuring that the output voltage of the DC side is stable within an allowable error range. Therefore, a double closed-loop control method is proposed for the 5L ANPC rectifier studied. The AC input control block diagram of the circuit topology is shown in Fig. 3.

1) Dual control loops: The dual closed-loop control strategy of the voltage outer loop and current inner loop is adopted. The outer loop voltage control mainly tracks the DC side voltage magnitude to realize voltage stability, so as to reduce the DC side voltage fluctuation as much as possible. By taking the difference between the DC side voltage sampled in real time and the reference voltage, the difference is processed and fed back to the system by the voltage regulator to control the DC voltage of the bus side. The current control target of the inner loop is the current magnitude of the filter inductor on the AC side. Compared with the given signal of the inner current loop obtained after the setting of the outer voltage loop, the power factor can be adjusted to keep the current sinusoidal. The signal processed by the current controller is used as the modulated signal of the main circuit switching device, which is compared with the triangular carrier to generate a PWM wave.

The switching function of the switches S_1 and S_3 is defined as S_{f13}, the switching function of switches S_5 and S_7 is respec-



PI: proportional-integral PR: proportion resonant

▲Figure 3. Feedforward control block diagram tively S_{f5} and S_{f7}, and then the output voltage can be written as:

$$V_o = \left[2 \cdot (S_{f13} - 1) + S_{f5} + S_{f7} \right] \times E \, . \tag{1}$$

$2E$ is selected as the voltage base value, and then the standardized output voltage range is [−1, 1]. To make the switches S_1 and S_3 work at the fundamental frequency, S_{f13} can be written as:

$$S_{f13} = \begin{cases} 1, 1 \leqslant v_o \leqslant 0 \\ 0, 1 \leqslant v_o \leqslant 0 \end{cases} , \tag{2}$$

where v_o is the reference value of the single-phase input voltage. Phase-shift pulse width modulation (PSPWM) is adopted in this paper with superior ability of FC voltage control. According to the principle of PSPWM modulation, the modulation voltage u_{ref} of the switches $S_5 - S_8$ can be written as:

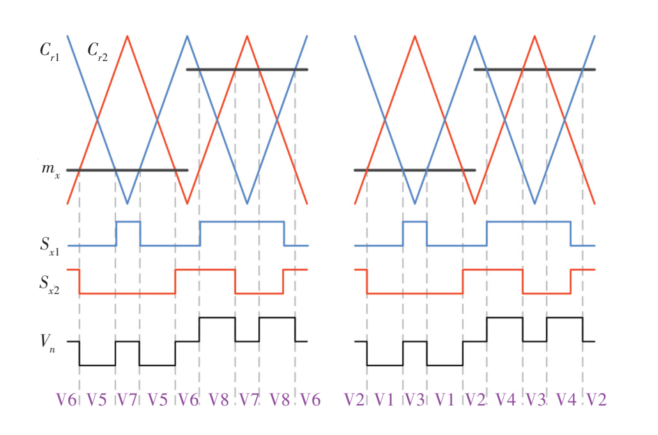
$$u_{ref} = \begin{cases} v_o, 0 \leqslant v_o \leqslant 1 \\ v_o + 1, -1 \leqslant v_o \leqslant 0 \end{cases} . \tag{3}$$

Under the SPWM modulation, the reference sinusoidal voltage of the single-phase output of the converter is:

$$u_o = m \cdot V_m \cdot \sin\theta \, , \tag{4}$$

where V_m is the AC voltage amplitude when the modulation ratio is maximum, $V_m = 2E = V_{dc}/2$, and V_{dc} is the DC side voltage; m represents the modulation ratio, where $m = u_m/V_m$, $0 \leqslant m \leqslant 1$, and u_m is the actual output voltage amplitude; θ represents the voltage phase angle, where $\theta = 2\pi ft$, and f is the voltage frequency.

2) FC voltage regulation: One of the major control difficulties of 5L ANPC is to balance the FC voltage to achieve the desired output voltage. As can be seen from Fig. 4, the balance of the voltage at both ends of C_f determines whether five levels can be generated normally. Therefore, ensuring the stability of the flying capacitor voltage is crucial to normal operation. If the voltage fluctuation is too large, the output voltage quality cannot be guaranteed, resulting in serious waveform distortion and other problems. On the other hand, the voltage fluctuation of the flying capacitor will directly affect the voltage stress of the switching devices. Under normal working conditions, the voltage borne by each switch is about $V_{dc}/4$, and the voltage fluctuation will directly lead to high voltage stress of the switches. Therefore, considering the above two considerations, the control goal must be achieved within the range of voltage stress constraints and output harmonic distortion constraints of switching devices.



▲Figure 4. Modulation diagram under different modulated wave m_x range

There is a 180° phase shift between carriers C_{r1} and C_{r2}, corresponding to S_{x1} and S_{x2}, respectively. As can be seen from the above section, the FC voltage can be adjusted by changing the working time of the redundant switching state (V_2, V_3) or (V_6, V_7) within one carrier cycle, which can be easily realized by modifying the modulated waves of S_{x1} and S_{x2}. For example, when the S_{x1} modulated wave residence time increases Δm_x and the S_{x2} modulated wave residence time decreases Δm_x, there is a total of $2\Delta m_x i_x$ current over one current-carrying cycle to charge the FC. Define the average terminal voltage of a carrier cycle as v_a and calculate it to satisfy the volt-second balance. At this time, the average current through FC can be written as:

$$\bar{i}_f = \frac{2\Delta t}{T_s} \times i_o \, . \tag{5}$$

The time width is very small and can be adjusted with a proportional-integral (PI) controller or a hysteresis comparator, or a constant value can be selected. With this method, the FC voltage can be easily stabilized near its reference value.

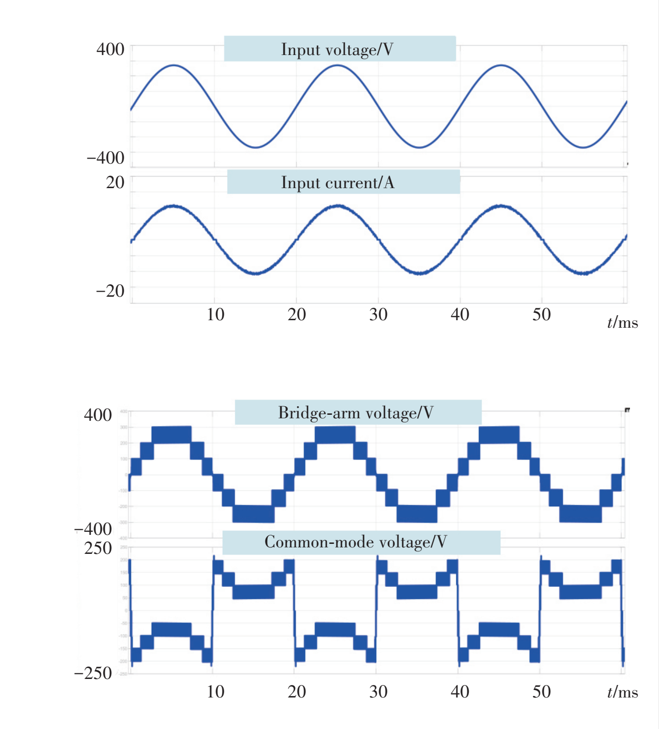
3 Simulation and Experimental Results

A 5L ANPC rectifier simulation model is established under the MATLAB/Simulink simulation tool, and the simulation analysis is carried out according to the required technical indicators.

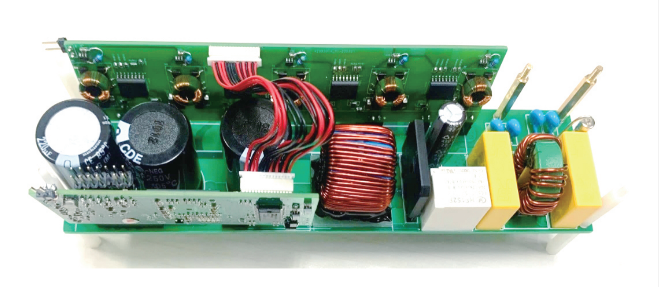
In an AC input condition, it can be seen from Fig. 5 that the bridge arm voltage has seven levels, and the common-mode voltage is small. At this time, the sinusoidal degree of the input current is good and basically consistent with the voltage phase. The power factor is 1, which can realize the unit power operation.

In order to verify the effectiveness of the topology and control strategy in this paper, a 1 kW experimental prototype is built as shown in Fig. 6. The input port is located on the right side of this figure while the output port on the left side. The two vertical circuit boards are control and drive circuits respectively.

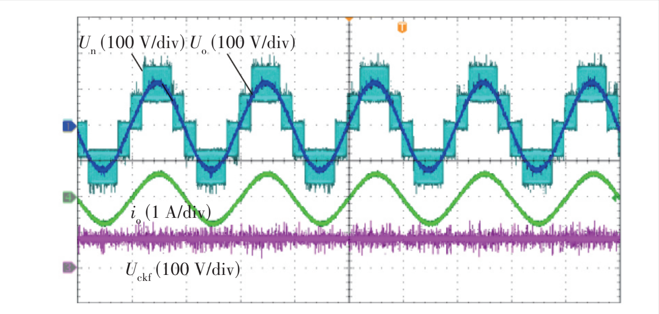
Firstly, the feasibility of the modulation strategy adopted in the reverse inverter experiment is verified. Fig. 7 shows the waveform of the bridge arm levels, the output AC voltage, the



▲**Figure 5. Waveform of alternating current (AC) input condition**



▲**Figure 6. Experimental prototype**

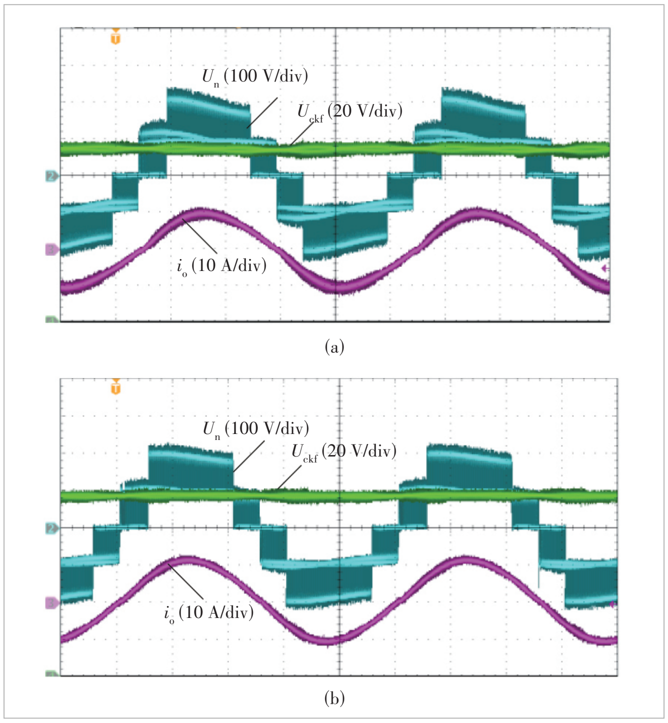


▲**Figure 7. Key waveforms of inversion**

output AC and the FC voltage. The DC bus voltage is 375 V and the load is 120 Ω. In this case, the output voltage is standard five levels, each level of which is $V_{dc}/4$, and the flying capacitor voltage is stable at $V_{dc}/4$. The output current has low harmonic content.

As shown in Fig. 8, the rectifier capability is verified. The input voltage is 220 V_{ac} , the peak input current is 10 A, the output voltage is 400 V_{dc}, and the FC voltage is 100 V. Fig. 8(a) shows that the voltage ripple across the FC is 12 V and the current ripple is about 1 A. As shown in Fig. 8(b), after adopting the FC voltage balance control, the FC voltage can basically maintain balanced near the reference value. The voltage ripple is 6 V and the current ripple is about 0.4 A. It can be seen that the level stability is improved, and THD is reduced from 6.72% to 4.27%, which verifies the effectiveness of the flying capacitor control strategy.

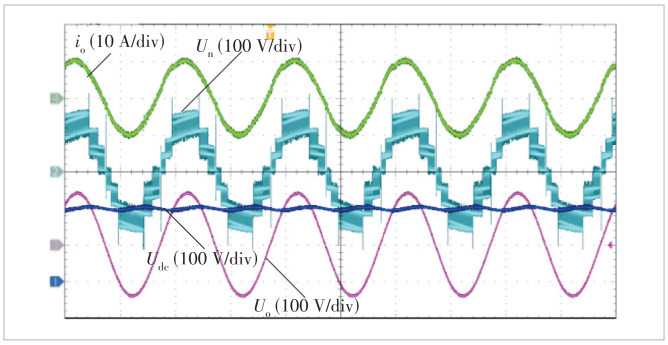
Fig. 9 shows the waveform of dual control loops rectifier condition, including input current i_o, terminal voltage V_n, output voltage V_{dc}, and input voltage V_{ac}. The input is 220 VAC and the load is 150 Ω. It can be seen that the combination of two bridge arms produces seven voltage levels. The second-



(a)

(b)

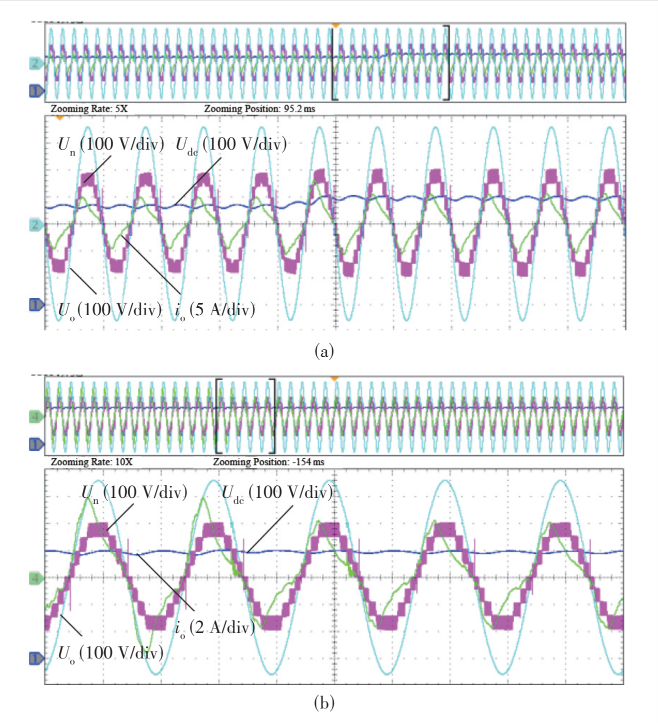
▲**Figure 8. Voltage and current waveforms of full load: (a) with open-loop control and (b) with flying-capacitor (FC) voltage control strategy**



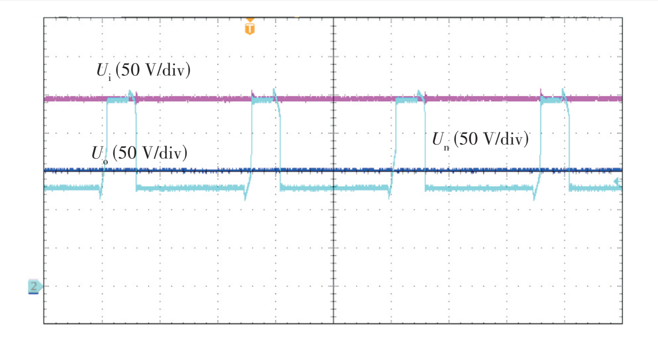
▲**Figure 9. Experimental waveforms of rectification under dual control loops** order voltage ripple of the output voltage is less than 20 V, and the current sinusoidal degree satisfies THD=3.89%.

Fig. 10 shows the experimental waveform of transient switching. As shown in Fig. 10(a), when the reference voltage changes from 370 V to 400 V, the output voltage reaches the new steady-state value within one period (13 ms) without significant overvoltage. As shown in Fig. 10(b), in the case of full load switching to half load, the amplitude of AC stabilizes at 1/2 of its original value after a short fluctuation. Transient experimental results show that the closed-loop control strategy has good dynamic response performance.

The experimental waveform under the DC input condition is shown in Fig. 11, where the input voltage is 120 V and the output voltage is 200 V. The bridge arm voltage can be composed of two levels, namely, 100 V and 200 V. Therefore, the feasibility of the control strategy is verified.



▲**Figure 10. Dynamic characteristic test waveforms: (a) reference voltage value changing from 370 V to 400 V and (b) full load switching to half load**



▲**Figure 11. Experimental waveform under a direct current (DC) input case**

4 Conclusions

To achieve low common-mode voltage and high power density, this paper proposes a multilevel PFC topology suitable for the communication power supply of 5G base stations. It is a hybrid topology consisting of a five-level ANPC bridge and a diode bridge. A 1 kW experimental prototype is established, which verifies the proposed working principle and control strategy.

Using the modulation strategy of PS-PWM, the converter produces seven levels while the d*v*/d*t* is reduced to $V_{dc}/7$, which greatly inhibits the common-mode voltage and reduces the leakage current. Moreover, by adjusting the redundancy switching state action time to balance the flying capacitor voltage, the total harmonic distortion can be suppressed by less than 4%. A dual closed loop system with PI control of the voltage outer loop and quasi-proportion resonant (quasi-PR) control of the current inner loop is used to realize zero static error tracking. The output voltage is stable within an allowable deviation range, and the control performance is great. Furthermore, the AC and DC input switching is realized to ensure the reliability and flexibility of the power supply.

References

[1] YAN X C, TENG H Y, PING L, et al. Study on security of 5G and satellite converged communication network [J]. ZTE communications, 2021, 19(4): 79 – 89. DOI: 10.12142/ZTECOM.202104009

[2] HOU X L, LI X, WANG X, etc. Some observations and thoughts about reconfigurable intelligent surface application for 5G evolution and 6G [J]. ZTE Communications, 2022, 20(1):14 – 20. DOI: 10.12142/ZTECOM.202201003

[3] WU Q Q, CHEN J Z, WU Z Q, et al. Synthesis and design of 5G duplexer based on optimization method [J]. ZTE communications, 2022, 20(3): 70 – 76. DOI: 10.12142/ZTECOM.202203009

[4] LUO F L. Signal processing techniques for 5G: an overview [J]. ZTE communications, 2015, 13(1): 20 – 27. DOI: 10.3969/j.issn.1673-5188.2015.01.003

[5] WU X Z, QI J J, LIU J D, et al. Review of multilevel inverter topology research using switched capacitor/switched inductor [J]. Proceedings of the CSEE, 2020, 40(1): 222 – 233, 389. DOI: 10.13334/j.0258-8013.pcsee.190323

[6] WANG K, ZHENG Z D, XU L, et al. An optimized carrier-based PWM method and voltage balancing control for five-level ANPC converters [J]. IEEE transactions on industrial electronics, 2020, 67(11): 9120 – 9132. DOI: 10.1109/TIE.2019.2956370

[7] YANG Y, PAN J Y, WEN H Q, et al. Double-vector model predictive control for single-phase five-level actively clamped converters [J]. IEEE transactions on transportation electrification, 2019, 5(4): 1202 – 1213. DOI: 10.1109/TTE.2019.2950510

[8] TAN G J, DENG Q W, LIU Z. An optimized SVPWM strategy for five-level active NPC (5L-ANPC) converter [J]. IEEE transactions on power electronics, 2013, 29(1): 386 – 395. DOI: 10.1109/TPEL.2013.2248172

[9] WANG K, XU L, ZHENG Z D, et al. Capacitor voltage balancing of a five-level ANPC converter using phase-shifted PWM [J]. IEEE transactions on power electronics, 2015, 30(3): 1147 – 1156. DOI: 10.1109/TPEL.2014.2320985

[10] ZHANG P, WU X Z, HE S, et al. A second-order voltage ripple suppression strategy of five-level flying capacitor rectifiers under unbalanced AC voltages

[J]. IEEE transactions on industrial electronics, 2023, 70(2): 1140 – 1149. DOI: 10.1109/TIE.2022.3159958

Biographies

TIAN Ruihan received her BS degree in electrical engineering and automation from Beijing Jiaotong University, China in 2021, where she is currently pursuing her MS degree in electrical engineering. Her current research interests include multilevel converters and DC/DC converters.

WU Xuezhi received his BS and MS degrees in electrical engineering from Beijing Jiaotong University, China in 1996 and 1999, respectively, and his PhD degree in electrical engineering from Tsinghua University, China in 2003. He is currently a professor with the School of Electrical Engineering, Beijing Jiaotong University. His current research interests include microgrids, wind power generation systems, power converters for renewable generation systems, power quality, and motor control.

XU Wenzheng (xuwenzheng@bjtu.edu.cn) received his BS degree in electrical engineering from Beijing Jiaotong University, China in 2012, MSc degree (with Distinction) in energy engineering from The University of Hong Kong, China in 2013, and PhD degree in electrical engineering from The Hong Kong Polytechnic University, China in 2020. He is currently a lecturer with the School of Electrical Engineering, Beijing Jiaotong University. His research interests include power electronics, wireless power transfer, transportation electrification, and energy storage converters.

ZUO Zhiling received his BS degree in automation from Hebei University of Science and Technology, China in 2002. He is the director of R&D Department of Power Platform in ZTE Corporation. He is now mainly engaged in communication power technology research and product development, mainly focusing on the architecture and development trend of communication power supply.

CHEN Changqing received his BS degree in electronic science and technology from Southwest Jiaotong University, China in 2006. He is currently working in ZTE Corporation, engaged in communication power supply technology research and product development. His research interests include high efficiency and high power density power supply, topology development, EMC and loss optimization and magnetic integration technology.

Mixed Electric and Magnetic Coupling Design Based on Coupling Matrix Extraction

XIONG Zhiang[1], ZHAO Ping[1], FAN Jiyuan[1],
WU Zengqiang[2], GONG Hongwei[2]
(1. Xidian University, Xi'an 710000, China;
2. ZTE Corporation, Shenzhen 518057, China)

DOI: 10.12142/ZTECOM.202304011
https://kns.cnki.net/kcms/detail/34.1294.TN.20231107.1357.002.html, published online November 8, 2023
Manuscript received: 2023-04-28

Abstract: This paper proposes a design and fine-tuning method for mixed electric and magnetic coupling filters. It derives the quantitative relationship between the coupling coefficients (electric and magnetic coupling, i. e., E_C and M_C) and the linear coefficients of frequency-dependent coupling for the first time. Different from the parameter extraction technique using the bandpass circuit model, the proposed approach explicitly relatesE_C and M_C to the coupling matrix model. This paper provides a general theoretic framework for computer-aided design and tuning of a mixed electric and magnetic coupling filter based on coupling matrices. An example of a 7th-order coaxial combline filter design is given in the paper, verifying the practical value of the approach.

Keywords: coupling matrix; frequency-dependent coupling; mixed electric and magnetic coupling; parameter extraction

Citation (Format 1): XIONG Z A, ZHAO P, FAN J Y, et al. Mixed electric and magnetic coupling design based on coupling matrix extraction [J]. ZTE Communications, 2023, 21(4): 85 – 90. DOI: 10.12142/ZTECOM.202304011
Citation (Format 2): Z. A. Xiong, P. Zhao, J. Y. Fan, et al., "Mixed electric and magnetic coupling design based on coupling matrix extraction," *ZTE Communications*, vol. 21, no. 4, pp. 85 – 90, Dec. 2023. doi: 10.12142/ZTECOM.202304011.

1 Introduction

In the design of microwave filters, the realization of finite transmission zeros (TZs) is critical to improving selectivity. Cross-coupling is the most popular method to create TZs[1]. However, this multi-path mechanism often leads to complexity in the design of filter layout, especially for high-order filters with many TZs. To solve this problem, frequency-dependent coupling (FDC) is introduced into the filter design. In an FDC, the coupling coefficient will be zero at a specific frequency, creating extra TZs in a given filter network.

SZYDLOWSKI et al. [2 – 5] proposed an optimization-based approach to the synthesis of coupling matrices with FDCs. Years later, HE[6 – 7] and ZHAO[8] developed deterministic matrix transformation approaches that can eliminate one cross-coupling from traditional *N*-tuples and introduce FDCs into the network. Constant couplings can be realized as pure electric or magnetic coupling, whereas FDCs need to be implemented as mixed electric and magnetic couplings. However, the mixed electric and magnetic coupling is difficult to control, because there is no quantitative relationship between FDCs and the mixed electric and magnetic couplings.

In 2006, MA[9] proposed constructing an electrical coupling and a magnetic coupling path between two resonators to generate a TZ. However, he did not give the relationship between the TZ position and the electric and magnetic coupling coefficients. In 2008, CHU[10 – 11] defined the mixed electric and magnetic coupling coefficient and gave the extraction method of the electric coupling coefficient (E_C) and the magnetic coupling coefficient (M_C) from the electromagnetic (EM) simulation of mixed electric and magnetic coupling structures. Furthermore, the relationship between the location of TZ and (E_C, M_C) is found. However, the parameter extraction is carried out in the bandpass domain. The approach does not explicitly relate Ec and Mc to the coupling matrix model, which is popular in filter synthesis. Therefore, it is difficult to design or tune the mixed coupling by coupling matrix extraction approaches.

This paper derives the explicit relationship between E_C (M_C) in the mixed electric and magnetic coupling and elements in the coupling and capacitance matrices. With the coupling matrix extracted by the model-based vector fitting (MVF) technique[12], the filter designer can easily design and tune the mixed coupling filter by comparing the extracted matrices with target ones.

The rest of the paper is organized as follows. Section 2 derives the relationship between the lowpass FDC model and the bandpass mixed coupling coefficients (E_C and M_C). Section 3 presents a mixed electric and magnetic coupling physical

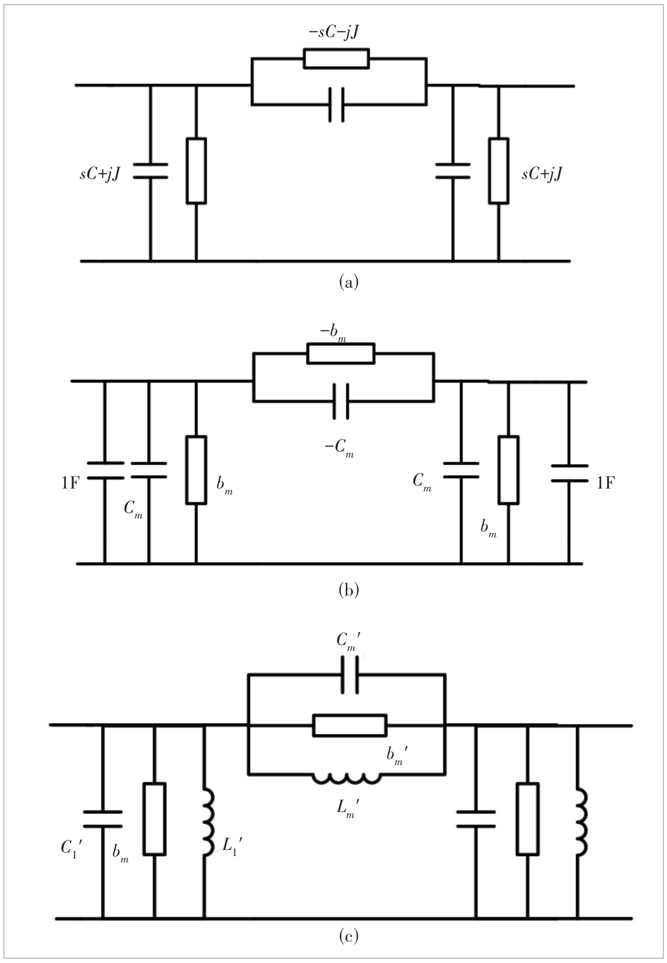
This work was supported by the National Natural Science Foundation of China under Grant No. 62001339.

model. The theory proposed in Section 2 is applied to design the mixed coupling structure. We then demonstrate a 7th-order in-line mixed coupling filter design with the aid of coupling matrix extraction. Section 4 concludes this paper.

2 Relationship Between FDC and Mixed Electric and Magnetic Coupling

A constant coupling in the coupling matrix is modeled by an ideal J-inverter, the π-equivalent circuit of which consists of three frequency-invariant susceptances (FISs). The characteristic admittance of a frequency-dependent inverter varies with frequency. Fig. 1 shows the π-equivalent circuit model of the frequency-dependent inverter. The circuit model includes three capacitors parallel-connected with FISs.

Note that the FDC is an element in lowpass circuit models. The bandpass frequency is mapped to the lowpass frequency domain by:



▲ **Figure 1. π - equivalent circuit model: (a) frequency-dependent inverter, where the admittances of capacitance and frequency-invariant susceptances are $sC(-sC)$ and $jJ(-jJ)$ respectively; (b) frequency-dependent inverter coupled lowpass network consisting of two resonators, where the two resonators are unit capacitors and resonant frequency is zero rad/s; (c) bandpass circuit model of mixed electric and magnetic coupling**

$$\Omega = \frac{\omega_0}{\omega_2 - \omega_1}\left(\frac{\omega}{\omega_0} - \frac{\omega_0}{\omega}\right), \qquad (1)$$

where Ω is the normalized lowpass frequency, ω_0 is the center frequency of the bandpass filter, ω_2 is the upper band edge frequency, ω_1 is the lower band edge, and ω is the bandpass frequency.

Substituting Eq. (1) into the admittance formula of FIS and capacitor connected in parallel yields

$$Y = jB + j\frac{\omega_0}{\omega_2 - \omega_1}\left(\frac{\omega}{\omega_0} - \frac{\omega_0}{\omega}\right)C_{\mathrm{LP}} = jB + j\omega C_{\mathrm{BP}} + \frac{1}{j\omega L_{\mathrm{BP}}}, \qquad (2)$$

where C_{LP} is the capacitance in the lowpass circuit model, C_{BP} is the capacitance, and L_{BP} is the inductance in the bandpass circuit model. According to Eq. (2), the parameters in the lowpass and bandpass circuits are related by:

$$C_{\mathrm{BP}} = \frac{C_{\mathrm{LP}}}{2\pi\mathrm{BW}}, L_{\mathrm{BP}} = \frac{2\pi\mathrm{BW}}{\omega_0^2 C_{\mathrm{LP}}}, \qquad (3)$$

where BW is the bandwidth and $2\pi\mathrm{BW}=\omega_2-\omega_1$. After transformation, the coupling between two resonators is not a pure electric or magnetic coupling form but mixed coupling. Therefore, an FDC should be realized as a mixed electric and magnetic coupling. However, Eq. (3) cannot reveal the qualitative relationship between FDC and mixed electric and magnetic coupling.

As shown in Fig. 1(b), there are unity capacitors on both sides of the frequency-dependent inverter. The capacitors model parallel resonant circuits with a resonant frequency of zero rad/s. The value of the frequency-dependent inverter is $sC_m + jb_m$. If the left node index is i and the right node index is j, the coupling matrix element M_{ij} is b_m, and the capacitance matrix element C_{ij} is C_m. After lowpass-to-bandpass circuit transformation in Eq. (3), the resultant bandpass circuit model is shown in Fig. 1(c), where

$$C_1' = \frac{1 + C_m}{2\pi\mathrm{BW}}, \quad C_m' = -\frac{C_m}{2\pi\mathrm{BW}}, \qquad (4a)$$

$$L_1' = \frac{2\pi\,\mathrm{BW}}{\omega_0^2\left(1 + C_m\right)}, \quad L_m' = -\frac{2\pi\mathrm{BW}}{\omega_0^2 C_m}, \qquad (4b)$$

$$b_m' = -b_m, \qquad (4c)$$

$$\omega_m \approx \omega_0 - \frac{b_m \pi\mathrm{BW}}{C_m}. \qquad (4d)$$

In the above derivation, we use the narrowband condition of ω_0 >> BW. This condition also applies to the derivation of Eqs. (7), (9) and (10). ω_m is the resonant frequency. When the frequency is ω_m, the parallel C_m', L_m' and b_m' form an open circuit, and the signal transmission is blocked to generate a TZ.

For a mixed electric and magnetic coupling, the calculation formula of Ec and Mc can be expressed as follows[10]:

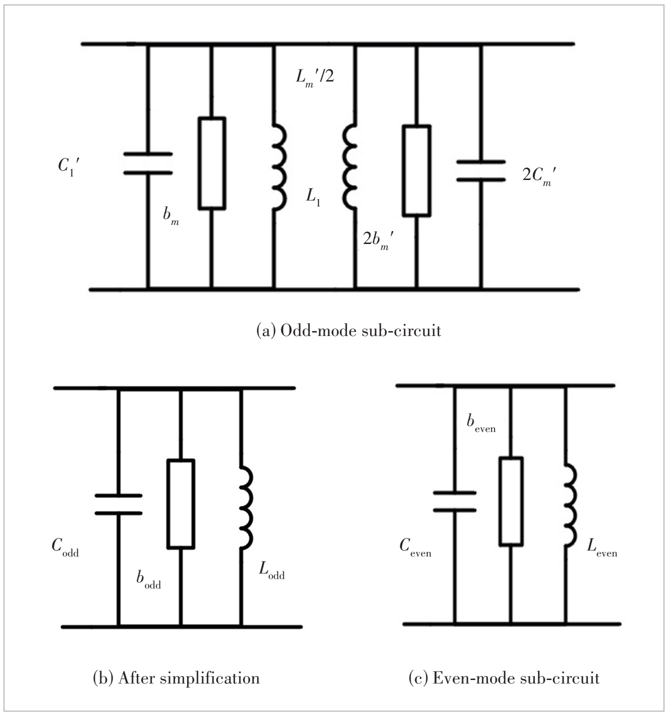
$$\mathrm{E_C} = \frac{{\omega_{\mathrm{ev}}}^2 - {\omega_{\mathrm{od}}}^2}{{\omega_{\mathrm{ev}}}^2 + {\omega_{\mathrm{od}}}^2 - 2{\omega_m}^2}, \tag{5a}$$

$$\mathrm{M_C} = \frac{{\omega_m}^2\left({\omega_{\mathrm{ev}}}^2 - {\omega_{\mathrm{od}}}^2\right)}{2{\omega_{\mathrm{ev}}}^2{\omega_{\mathrm{od}}}^2 - {\omega_m}^2\left({\omega_{\mathrm{ev}}}^2 + {\omega_{\mathrm{od}}}^2\right)^2}, \tag{5b}$$

where ω_{ev} is the even mode resonant frequency, and ω_{od} is the odd mode resonant frequency of a coupled resonator pair.

We can calculate $\mathrm{E_C}$ and $\mathrm{M_C}$ in the mixed electric and magnetic coupling based on ω_m, ω_{ev}, and ω_{od}. To obtain $\mathrm{E_C}$ and $\mathrm{M_C}$ of the mixed electric and magnetic coupling in terms of FDC coefficients, we can analyze the even- and odd-mode resonant frequencies of the coupled-resonator circuit model in Fig. 1(c).

We analyze the odd mode first. The odd-mode sub-circuit is shown in Fig. 2(a). After combining parallel-connected capacitors, inductors, and FISs, the odd-mode sub-circuit is transformed into the form shown in Fig. 2(b). We have:



(a) Odd-mode sub-circuit

(b) After simplification

(c) Even-mode sub-circuit

▲Figure 2. Mode circuit of a bandpass circuit model of mixed electric and magnetic coupling

$$C_{\mathrm{odd}} = C_1' - 2C_m' = \frac{1 - C_m}{2\pi\mathrm{BW}}, \tag{6a}$$

$$L_{\mathrm{odd}} = \frac{2C_1'C_m'}{C_1' + 2C_m'} = \frac{2\pi\mathrm{BW}}{\omega_0^2 - \omega_0^2 C_m}, \tag{6b}$$

$$b_{\mathrm{odd}} = -b_m. \tag{6c}$$

Therefore, the resonant frequency of the odd mode is

$$\omega_{\mathrm{od}} = \frac{\sqrt{b_{\mathrm{odd}}^2 + 4\dfrac{C_{\mathrm{odd}}}{L_{\mathrm{odd}}}} - b_{\mathrm{odd}}}{2C_{\mathrm{odd}}} \approx \omega_0 + \frac{b_m\pi\mathrm{BW}}{1 - C_m}. \tag{7}$$

Similarly, to analyze the even mode, as shown in Fig. 2(c), we have

$$C_{\mathrm{even}} = \frac{1 + C_m}{2\pi\mathrm{BW}}, \tag{8a}$$

$$L_{\mathrm{even}} = \frac{2\pi\mathrm{BW}}{\omega_0^2\left(1 + C_m\right)}, \tag{8b}$$

$$b_{\mathrm{even}} = b_m. \tag{8c}$$

Therefore, the resonant frequency of the even mode sub-circuit is

$$\omega_{\mathrm{even}} = \frac{\sqrt{b_{\mathrm{even}}^2 + 4\dfrac{C_{\mathrm{even}}}{L_{\mathrm{even}}}} - b_{\mathrm{even}}}{2C_{\mathrm{even}}} \approx \omega_0 - \frac{b_m\pi\mathrm{BW}}{1 + C_m}. \tag{9}$$

Substituting Eqs. (7) and (9) into Eq. (5) yields

$$\mathrm{E_C} = \frac{{\omega_{\mathrm{ev}}}^2 - {\omega_{\mathrm{od}}}^2}{{\omega_{\mathrm{ev}}}^2 + {\omega_{\mathrm{od}}}^2 - 2{\omega_m}^2} \approx -C_m, \tag{10a}$$

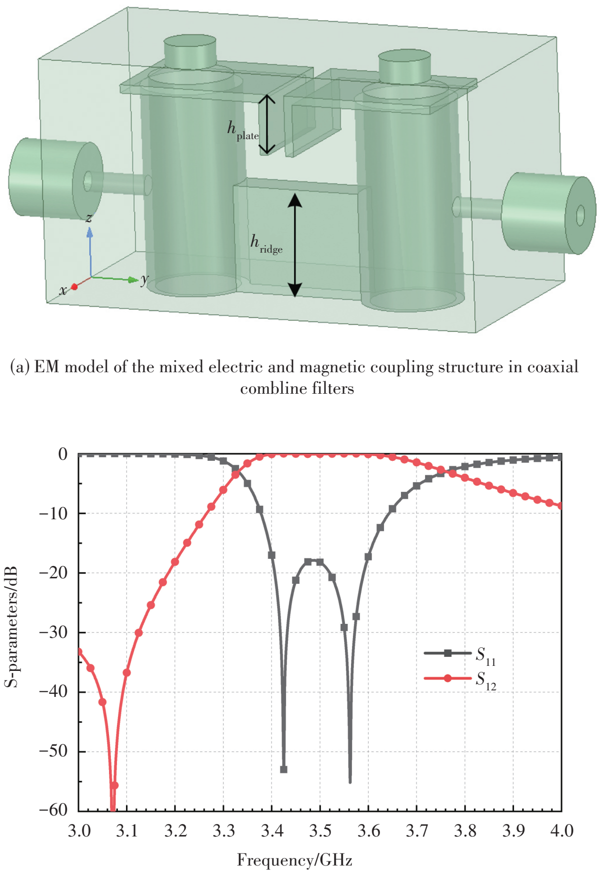
$$\mathrm{M_C} = \frac{{\omega_m}^2\left({\omega_{\mathrm{ev}}}^2 - {\omega_{\mathrm{od}}}^2\right)}{2{\omega_{\mathrm{ev}}}^2{\omega_{\mathrm{od}}}^2 - {\omega_m}^2\left({\omega_{\mathrm{ev}}}^2 + {\omega_{\mathrm{od}}}^2\right)^2} \approx -C_m. \tag{10b}$$

The results in Eq. (10) show that $\mathrm{E_C}$ and $\mathrm{M_C}$ in mixed electric and magnetic coupling filters are almost equal, and both values are approximately equal to $-C_m$. This analysis result reveals that the majority of electric and magnetic coupling should be canceled with each other to realize an FDC. The electric coupling or magnetic coupling is slightly stronger than the other one to provide a weak total coupling for constructing the narrowband passband. Therefore, if the absolute value of the synthesized capacitance matrix element C_{ij} is larger, the electric and magnetic coupling in the mixed coupling structure should be tuned stronger simultaneously.

To conclude, the FDC in the lowpass coupling matrix model is $sC_{ij}+jM_{ij}$, where M_{ij} represents the total coupling exhibited by the mixed electric and magnetic coupling at the center frequency, and C_m is related to the strength of both the electric and the magnetic coupling coefficient in the mixed coupling.

3 Analysis of Electromagnetic Model

For the experimental validation, a 7th-order in-line bandpass filter is designed with coaxial cavity structures in this section. The 7th-order filter contains two mixed electric and magnetic couplings, the structure of which is shown in Fig. 3(a). The simulation results of the second-order filter block are shown in Fig. 3(b). The center frequency of the filter is 3.5 GHz, the bandwidth is 0.2 GHz, and the return loss is 18 dB. The open end of the metal rod is connected to a folded metal sheet. Two adjacent metal sheets form a parallel plate capacitor to realize a strong electric coupling. The height of the plate h_{plate} is 3.6 mm. The short ends of adjacent coaxial resonators are connected by a metal ridge to realize a strong magnetic coupling. The height of the ridge, h_{ridge}, is 6.3 mm. The strong electrical coupling and the magnetic coupling exist simultaneously to form a mixed electric and magnetic coupling. If h_{plate} or h_{ridge} increases, the electric coupling or magnetic coupling will become stronger in this design.



(a) EM model of the mixed electric and magnetic coupling structure in coaxial combline filters

(b) Simulated response of the second-order filter

EM: electromagnetic

▲Figure 3. EM model and response of second-order filter

By repeatedly applying the MVF technique to extract the coupling matrix from simulation data[12], we can study the relationship between the mixed coupling coefficients and the element values of the coupling and capacitance matrices. Tables 1 and 2 show the extracted values of M_{ij} and C_{ij} when $\mathrm{E_C}$ and $\mathrm{M_C}$ are changed. It can be found from Table 1 that when h_{ridge} and h_{plate} increase simultaneously, C_{ij} increases, whereas M_{ij} almost does not change. Therefore, C_{ij} is related to both the electric and magnetic coupling coefficients in the mixed coupling.

Table 2 shows that when $\mathrm{E_C}$ increases and $\mathrm{M_C}$ decreases, M_{ij} increases. Since M_{ij} represents the total coupling exhibited by the mixed electric and magnetic coupling at the center frequency, it can be seen that $\mathrm{E_C}$ is stronger than $\mathrm{M_C}$ in the mixed coupling structure shown in Fig. 3(a). Table 2 also shows that we can control M_{ij} by increasing the difference between $\mathrm{E_C}$ and $\mathrm{M_C}$ without affecting C_{ij}.

Table 3 shows that when $\mathrm{M_C}$ increases, the TZ is shifted to

▼Table 1. Simultaneously changing the heights of the ridge and the plate

h_{ridge}/mm	h_{plate}/mm	M_{ij}	C_{ij}	TZ/rad
6.70	3.96	1.436 2	0.369 5	−3.89
6.50	3.78	1.440 9	0.355 3	−4.06
6.30	3.60	1.440 2	0.343 3	−4.20
6.10	3.43	1.443 8	0.331 8	−4.35
5.90	3.27	1.438 3	0.320 9	−4.48

TZ: transmission zero

▼Table 2. Changing the height of the ridge or the plate

h_{ridge}/mm	h_{plate}/mm	M_{ij}	C_{ij}	TZ/rad
6.70	3.43	0.794 5	0.341 8	−2.32
6.50	3.50	1.097 2	0.341 8	−3.21
6.30	3.60	1.440 2	0.343 3	−4.20
6.10	3.68	1.761 0	0.343 7	−5.12
5.90	3.75	2.056 8	0.343 8	−5.98

TZ: transmission zero

▼Table 3. Changing the height of the ridge or the plate when TZ is in the lower stopband

h_{ridge}/mm	h_{plate}/mm	M_{ij}	C_{ij}	TZ/rad
6.30	3.60	1.440 2	0.343 3	−4.20
6.90	3.60	0.760 1	0.353 5	−2.15
7.50	3.60	−0.029 4	0.363 0	0.08
8.10	3.60	−0.912 7	0.372 8	2.45
8.70	3.60	−1.601 0	0.397 9	4.02

TZ: transmission zero

the right. From Table 4, it can be found that when E_C increases, the TZ is shifted to the left. From Tables 3 and 4, we can also see that when the TZ is located in the lower stopband, E_C is stronger than M_C. If the TZ is in the upper stopband, then M_C is stronger than E_C.

To conclude, the tuning of the mixed electric and magnetic coupling structure in Fig. 3(a) follows two rules:

Rule 1: We simultaneously increase or decrease E_C and M_C to tune C_{ij}.

Rule 2: If the TZ is in the lower stopband, we can increase E_C and decrease M_C to increase M_{ij}. If the TZ is in the upper stopband, we can increase M_C and decrease E_C simultaneously to increase M_{ij}.

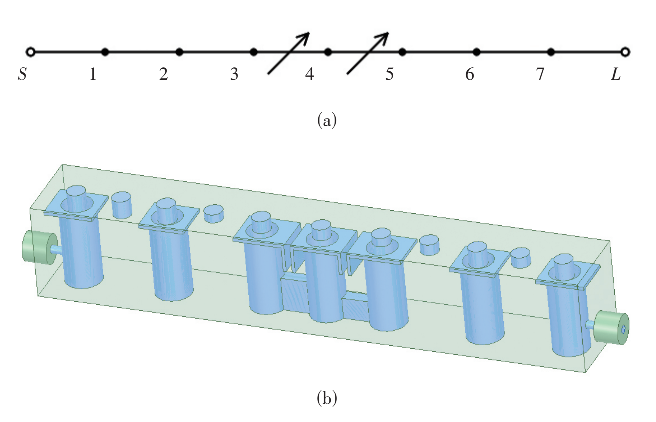
To verify the above theory, take a 7th-order filter with the coupling topology shown in Fig. 4(a) as an example. The center frequency and bandwidth of the filter are 3.5 GHz and 200 MHz, respectively. The in-band return loss level is required to be 18 dB. Two TZs at 3.7 GHz and 3.3 GHz are generated sequentially by two mainline FDCs. The synthesized coupling matrix and capacitance matrix are shown in Figs. 5 and 6, respectively.

The perspective view of the filter model is shown in Fig. 4(b). With the help of the MVF method to extract the coupling matrix from simulation data, we can identify how to adjust the dimensions and finally obtain satisfactory filter responses. The simulation results with ideal lossless materials are shown in Fig. 7, where solid lines are simulation data, and dashed lines are the ideal synthesis responses.

▼Table 4. Changing the height of the ridge or the plate when TZ is in the upper stopband

h_{ridge}/mm	h_{plate}/mm	M_{ij}	C_{ij}	TZ/rad
6.30	1.60	−0.958 2	0.236 2	4.06
6.30	2.10	−0.376 8	0.265 4	1.42
6.30	2.60	0.212 7	0.293 0	−0.73
6.30	3.10	0.822 3	0.318 9	−2.58
6.30	3.60	1.440 2	0.343 3	−4.20

TZ: transmission zero



(a)

(b)

▲Figure 4. 7th-order filter with mixed electric and magnetic coupling: (a) target topology and (b) electromagnetic model of the 7th-order filter with mixed electric and magnetic coupling

4 Conclusions

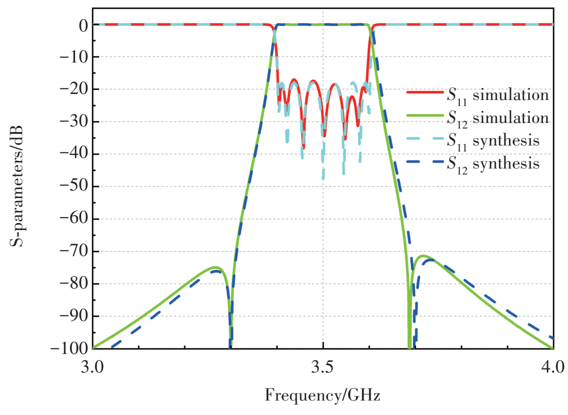
In this paper, the relationship between the coupling matrix (capacitance matrix) and E_C (M_C) is obtained through circuit analysis. A filter example is designed to verify the proposed theory. Although only the inline filter is discussed in detail, the strategy introduced in this paper can be easily generalized to mixed electric and magnetic coupling filters in different coupling topologies. Compared with the existing theory of mixed electric and magnetic coupling filters, this work has the following distinctive features.

0	0.952 2	0	0	0	0	0	0	0
0.952 2	0.000 4	0.793 8	0	0	0	0	0	0
0	0.793 8	0.000 5	0.555 8	0	0	0	0	0
0	0	0.555 8	0.342 2	0.580 9	0	0	0	0
0	0	0	0.580 9	0.023 1	0.571 4	0	0	0
0	0	0	0	0.571 4	0.321 4	0.559 5	0	0
0	0	0	0	0	0.559 5	0.000 5	0.793 8	0
0	0	0	0	0	0	0.793 8	0.000 4	0.952 2
0	0	0	0	0	0	0	0.952 2	0

▲Figure 5. Coupling matrix

0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0
0	0	0	1	−0.298 5	0	0	0	0
0	0	0	−0.298 5	1	0.277 3	0	0	0
0	0	0	0	0.277 3	1	0	0	0
0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0

▲Figure 6. Capacitance matrix



▲Figure 7. Response of the 7th-order, where dashed lines are ideal synthesis responses and solid lines are simulation results

1) It derives the explicit relationship between E_C(Mc) in the mixed electric and magnetic coupling and elements in the coupling and capacitance matrices.

2) The filter tuning procedure is based on analytical coupling matrix extraction and thus is very fast, compared with optimization-based filter tuning techniques.

This paper gives a guiding idea for designing the physical model of the mixed electric and magnetic coupling filter.

References

[1] CAMERON R J, KUDSIA C M, MANSOUR R R. Microwave filters for communication systems [M]. New York, USA: Wiley, 2018. DOI: 10.1002/9781119292371

[2] SZYDLOWSKI L, LAMECKI A, MROZOWSKI M. Coupled-resonator waveguide filter in quadruplet topology with frequency-dependent coupling – A design based on coupling matrix [J]. IEEE microwave and wireless components letters, 2012, 22(11): 553 – 555. DOI: 10.1109/LMWC.2012.2225604

[3] SZYDLOWSKI L, LESZCZYNSKA N, LAMECKI A, et al. A substrate integrated waveguide (SIW) bandpass filter in a box configuration with frequency-dependent coupling [J]. IEEE microwave and wireless components letters, 2012, 22(11): 556 – 558. DOI: 10.1109/LMWC.2012.2221690

[4] SZYDLOWSKI L, MROZOWSKI M. A self-equalized waveguide filter with frequency-dependent (resonant) couplings [J]. IEEE microwave and wireless components letters, 2014, 24(11): 769 – 771. DOI: 10.1109/LMWC.2014.2303171

[5] SZYDLOWSKI L, LESZCZYNSKA N, LAMECKI A, et al. A substrate integrated waveguide (SIW) bandpass filter in A box configuration with frequency-dependent coupling [J]. IEEE microwave and wireless components letters, 2012, 22(11): 556 – 558. DOI: 10.1109/LMWC.2012.2221690

[6] HE Y X, MACCHIARELLA G, WANG G, et al. A direct matrix synthesis for in-line filters with transmission zeros generated by frequency-variant couplings [J]. IEEE transactions on microwave theory and techniques, 2018, 66(4): 1780 – 1789. DOI: 10.1109/tmtt.2018.2791940

[7] HE Y X, WANG G, SUN L G, et al. Direct matrix synthesis for in-line filters with transmission zeros generated by frequency-variant couplings [C]//IEEE MTT-S International Microwave Symposium (IMS). IEEE, 2017: 356 – 359. DOI: 10.1109/MWSYM.2017.8059119

[8] ZHAO P, WU K. Cascading fundamental building blocks with frequency-dependent couplings in microwave filters [J]. IEEE transactions on microwave theory and techniques, 2019, 67(4): 1432 – 1440. DOI: 10.1109/TMTT.2019.2895532

[9] MA K X, MA J G, YEO K S, et al. A compact size coupling controllable filter with separate electric and magnetic coupling paths [J]. IEEE transactions on microwave theory and techniques, 2006, 54(3): 1113 – 1119. DOI: 10.1109/TMTT.2005.864118

[10] CHU Q X, WANG H. An inline coaxial quasi-elliptic filter with controllable mixed electric and magnetic coupling [J]. IEEE transactions on microwave theory and techniques, 2009, 57(3): 667 – 673. DOI: 10.1109/TMTT.2009.2013290

[11] CHU Q X, WANG H. A compact open-loop filter with mixed electric and magnetic coupling [J]. IEEE transactions on microwave theory and techniques, 2008, 56(2): 431 – 439. DOI: 10.1109/TMTT.2007.914642

[12] ZHAO P, WU K L. Model-based vector-fitting method for circuit model extraction of coupled-resonator diplexers [J]. IEEE transactions on microwave theory and techniques, 2016, 64(6): 1787 – 1797. DOI: 10.1109/TMTT.2016.2558639

Biographies

XIONG Zhiang received his BS degree from Jiangxi University of Science and Technology, China in 2021. He is currently pursuing his master's degree in electromagnetic wave and microwave technology at Xidian University, China. His current research interests include mixed electric and magnetic coupling filter, and modeling and optimization of microwave passive filters.

ZHAO Ping (aoing56@gmail.com) received his BS degree from Nanjing University, China, in 2012, and PhD degree from The Chinese University of Hong Kong, China in 2017. From 2017 to 2019, he was a post-doctoral researcher with the École Polytechnique de Montréal, Canada. In 2020, he joined the National Key Laboratory of Antennas and Microwave Technology, School of Electronic Engineering, Xidian University, China, as an associate professor. His research interests include coupling matrix synthesis techniques for coupled-resonator networks, analytical computer-aided tuning (CAT) algorithms for microwave and millimeter-wave filters, and diplexers with applications in cellular base stations and satellites. He is also interested in modeling and optimization of passive RF components and computer-aided design techniques, such as the homotopy method, artificial neural networks, and machine learning techniques. He is a member of IEEE.

FAN Jiyuan received his BS degree from the Nanjing University of Posts and Telecommunications, China in 2022. He is currently pursuing the master's degree in electro-magnetic wave and microwave technology at Xidian University, China.

WU Zengqiang received his master's degree from Xidian Univesity, China in 2018. The same year, he joined ZTE Corporation as an assistant engineer. His current research interests include miniaturization techniques and lossless techniques of microwave filters for wireless communication, such as coaxial filter and dielectric waveguide filter.

Gong Hongwei joined ZTE Corporation as a senior engineer. His current research interests include miniaturization techniques and lossless techniques of microwave filters for wireless communication, such as the coaxial filter and the dielectric waveguide filter.

Beyond Video Quality: Evaluation of Spatial Presence in 360-Degree Videos

ZOU Wenjie[1], GU Chengming[1], FAN Jiawei[1], HUANG Cheng[2,3], BAI Yaxian[2,3]

(1. School of Telecommunications Engineering, Xidian University, Xi'an 710071, China;
2. ZTE Corporation, Shenzhen 518057, China;
3. State Key Laboratory of Mobile Network and Mobile Multimedia Technology, Shenzhen 518055, China)

DOI: 10.12142/ZTECOM.202304012

https: //kns.cnki.net/kcms/detail/34.1294.TN.20231129.1000.002.html, published online November 29, 2023

Manuscript received: 2023-05-08

Abstract: With the rapid development of immersive multimedia technologies, 360-degree video services have quickly gained popularity and how to ensure sufficient spatial presence of end users when viewing 360-degree videos becomes a new challenge. In this regard, accurately acquiring users' sense of spatial presence is of fundamental importance for video service providers to improve their service quality. Unfortunately, there is no efficient evaluation model so far for measuring the sense of spatial presence for 360-degree videos. In this paper, we first design an assessment framework to clarify the influencing factors of spatial presence. Related parameters of 360-degree videos and head-mounted display devices are both considered in this framework. Well-designed subjective experiments are then conducted to investigate the impact of various influencing factors on the sense of presence. Based on the subjective ratings, we propose a spatial presence assessment model that can be easily deployed in 360-degree video applications. To the best of our knowledge, this is the first attempt in literature to establish a quantitative spatial presence assessment model by using technical parameters that are easily extracted. Experimental results demonstrate that the proposed model can reliably predict the sense of spatial presence.

Keywords: virtual reality; quality assessment; omnidirectional video; spatial presence

Citation (Format 1): ZOU W J, GU C M, FAN J W, et al. Beyond video quality: evaluation of spatial presence in 360-degree videos [J]. ZTE Communications, 2023, 21(4): 91 – 103. DOI: 10.12142/ZTECOM.202304012

Citation (Format 2): W. J. Zou, C. M. Gu, J. W. Fan, et al., "Beyond video quality: evaluation of spatial presence in 360-degree videos," *ZTE Communications*, vol. 21, no. 4, pp. 91 – 103, Dec. 2023. doi: 10.12142/ZTECOM.202304012.

1 Introduction

In the past decade, multimedia streaming services have had an explosive growth[1]. Among a variety of multimedia types, 360-degree videos become the major type of virtual reality (VR) content in the current stage. Major video-sharing websites such as YouTube and Facebook have already started to offer 360-degree video-on-demand and live 360-degree video streaming services.

In contrast to traditional 2D videos, 360-degree videos can provide full 360-degree scenes to end users, using the Head-Mounted Display (HMD) as a display device. With a higher degree of freedom (DoF) and wider field of view (FOV) during the viewing process, end users are provided with a stronger sense of immersion and a feeling of being in a perceptible virtual scene around the users. Different from the experience of traditional 2D videos[2 – 3], this type of feeling is usually termed as presence[4 – 7]. According to the classification of presence in Refs. [8] and [9], presence covers a broad range of aspects including spatial presence, social presence, self-presence[10], engagement, realism, and cultural presence. In the field of 360-degree video processing, researchers are more interested in spatial presence, which describes the feeling, sense, or state of "being there" in a mediated environment[4]. This feeling occurs when part or all of a person's perception fails to accurately acknowledge the role of technology that makes it appear that she/he is in a physical location and environment different from her/his actual location and environment in the physical world[11].

Over the last twenty years, a variety of work has been conducted to investigate the users' sense of presence in VR environments, especially for scenes rendered by computers[12 – 13]. These studies mainly focused on measuring specific influencing factors of the sense of presence and revealing the qualitative relationship between presence and specific human perceptual aspects in a generalized VR environment. Directly quantifying the sense of presence is, however, outside the scope of

This work is supported in part by ZTE Industry-University-Institute Cooperation Funds.

these studies. On the other hand, some researchers managed to evaluate the sense of presence using physiological signals[14 - 17]. However, this type of method requires professional equipment and the reliability of experimental results strongly relies on the accuracy of the devices.

To the best of our knowledge, most human perception research carried out for 360-degree videos only focused on the perceptual video quality instead of the spatial presence. Recently, we conducted a subjective evaluation experiment on the spatial presence of end users when watching 360-degree videos displayed on VR devices[18]. We aimed to quantitatively investigate the relationship between various impact factors and the spatial presence.

In this paper, based on the research outcomes of Ref. [18], the characteristics of the display device of 360-degree videos are considered. We propose a framework in hierarchical structure to clarify the influencing factors of the spatial presence, where both the features of 360-degree video and HMD are considered. A series of rigorous subjective experiments are designed to reveal the relationship between various influencing factors and the spatial presence. Furthermore, a quantitative evaluation model of spatial presence is built in this work. Contributions of this paper can be concluded as follows:

1) We propose the first framework to identify the components of spatial presence. This framework provides valuable input for establishing models of assessing the spatial presence of VR services.

2) We reveal the relationship between spatial presence and various related impact factors based on subjective ratings, which can be used as recommendations for further improving the quality of 360-degree video services.

3) We propose the first quantitative model to measure the spatial presence when watching 360-degree videos on the HMD. The parameters employed in the proposed model can be easily extracted, hence the model would be conveniently deployed on the network or client to assess the user's presence.

The rest of this paper is organized as follows. Section 2 introduces the related work. Section 3 illustrates the assessment framework and the subjective experiments. Section 4 introduces the proposed model in detail. In Section 5, the performance of the proposed model is evaluated. Conclusions are drawn in Section 6.

2 Related Work

Over the last thirty years, researchers have explained and defined the concept of presence in several different ways. For instance, LOMBARD et. al.[8] defined it as the experience of being engaged by the representations of a virtual world in 2002. Very recently, presence was defined as the feeling of being in a perceptible external world around the self[4 - 7]. The evolution of understanding and definition of the presence was summarized in Refs. [7] and [9]. As the above research is more related to psychoanalysis, straightforward solutions to the measurement of presence were outside the scope of these studies. How to measure the presence in practice is still unknown.

To acquire the subjective sense of presence, some researchers resorted to the design of subjective response questionnaires[19 - 24]. More specifically, authors in Ref. [19] designed a questionnaire, called the immersive tendencies questionnaire (ITQ), to investigate the relationship between users' sense of presence and some handcrafted influential aspects in virtual environments. Authors in Ref. [22] designed a spatial presence questionnaire, named MEC Spatial Presence Questionnaire (MSC-SPQ), to investigate the influence of possible actions, self-location, and attention allocation on users' sense of spatial presence. However, these studies only focused on revealing the qualitative relationship between specific human perceptual aspects and presence in the generalized VR environment. On the other hand, some researchers tried to evaluate the presence using physiological signals[14 - 17]. This type of measurement requires the deployment of professional equipment which is impractical for real-world applications. Therefore, designing accurate and implementation-friendly experimental methods to measure presence is of fundamental importance.

As for the human perception research specifically carried out for 360-degree videos, to our best knowledge, most studies only focused on evaluating the quality of experience aspects[25 - 33] instead of assessing the sense of presence. For instance, authors in Ref. [25] investigated how to assess the video quality of 360-degree videos corresponding to different projection approaches. A quality metric, called spherical peak signal to noise ratio (S-PSNR) was proposed to summarize the average quality over all possible viewports as the video quality. In Ref. [26], authors proposed an objective video quality assessment method using a weighted PSNR and special zero area distortion projection method for 360-degree videos. In Ref. [30], authors measured viewport PSNR values over time to assess the objective video quality of 360-degree video streaming. Recently, authors in Ref. [33] introduced visual attention in assessing the objective quality of 360-degree videos with the assumption that not all of the 360-degree scene is actually watched by users. However, as discussed above, the spatial presence of end users was not fully considered in existing research. Our recent work[18] conducted a preliminary experiment for assessing the spatial presence of end users when viewing 360-degree videos displayed on VR devices. However, modeling the spatial presence of end users is not covered. How to quantitatively evaluate users' sense of spatial presence when viewing 360-degree videos remains an open issue.

3 Subjective Evaluation Framework and Subjective Experiments

In this section, a hierarchical framework with five perception modules is first proposed to assess spatial presence. Based on this framework, five subjective experiments were designed and conducted according to each module in the framework. Results of subjective experiments are used to investigate each type of human perception and facilitate the establishment of the assessment model.

3.1 Proposed Assessment Framework

As shown in Fig. 1, the proposed framework consists of three layers, namely the factor layer, the perception layer, and the presence layer from left to right. The factor layer includes several sensory cues of relevant parameters such as video, audio, VR device, and latency. These parameters can be conveniently extracted from the current VR systems. In the perception layer, users' perception is characterized into multiple dimensions including visual[34] , auditory[34], and interactive perception[21 – 22, 35]. Detailed definitions of the components of perception and presence layers are discussed as follows.

1) Perceptual video quality

Perceptual video quality refers to the overall perceived quality of videos displayed on the HMD. In our previous work, three technological parameters of the video, i.e., video bitrate, video resolution and video frame rate, are extracted to assess the video quality. Two parameters corresponding to the HMD (screen resolution and refresh rate) are added in the assessment of perceptual video quality.

2) Perceptual audio quality

Perceptual audio quality refers to the overall perceived quality of audios offered by the VR system. The audio bitrate and audio sampling rate are extracted to assess the perceptual audio quality.

3) Visual realism

Visual realism (VRE) refers to how close the system's visual output is to real-world visual stimuli. This perception not only depends on the video quality, but also depends on how wide the FOV provided by HMD is and whether a stereoscopic vision is offered. These two additional factors have been verified to be important for the overall capability of an immersive system[36].

4) Acoustic realism

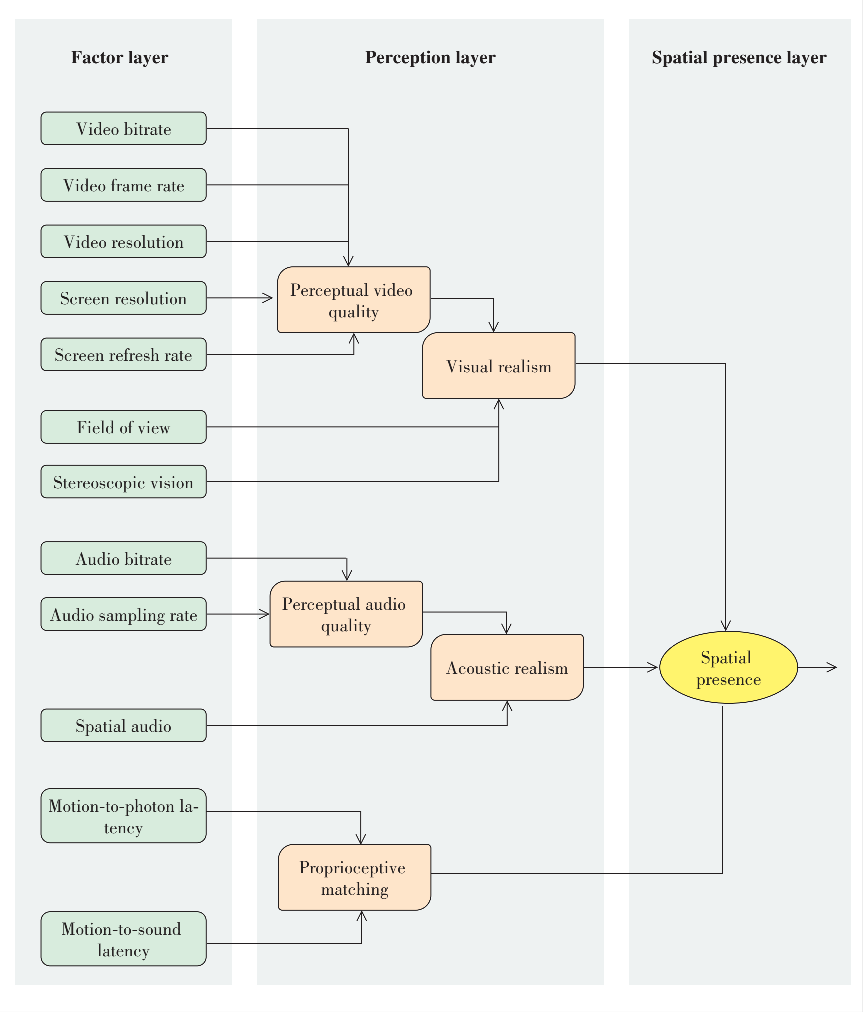
Acoustic realism (ARE) represents how close the system's aural output is to real-world aural stimuli. Perceptual audio quality is the basic experience of the audio. Moreover, spatial audio provides the capability to track sound directions and update the head movement in real time. Hence, the spatial audio and perceptual audio quality are combined to assess the overall acoustic realism.

5) Proprioceptive matching

Proprioceptive matching refers to the matching degree between the head movement and the picture/sound refresh of the HMD. As for a VR system, the tracking level is much more important in regard to the spatial presence formation[36]. Similarly, the mismatch can also occur in the spatial audio. These two mismatches, called motion-to-photon (MTP) latency and audio latency (AL)[37], are utilized to assess the capability of proprioceptive matching.

6) Spatial presence

Spatial presence refers to a user's subjective psychological response to a VR system[35]. It is correlated with VRE, ARE, and proprioceptive matching, which represents the main aspects of the experience provided by 360-degree video services.



▲Figure 1. Proposed assessment framework for assessing spatial presence

3.2 Subjective Experiments for Obtaining Spatial Presence

To explore the spatial presence, six subjective quality scoring experiments were conducted, corresponding to the five perception modules in the perception layer and one towards the spatial presence.

3.2.1 Overview of Experimental Design

A total number of 30 non-expert subjects participated in this experiment, including

16 males and 14 females aged between 22 and 33 years. All of them have normal or corrected-to-normal sight. The experiments were conducted in the test environment following ITU-T P.913[38]. A flagship HMD, i.e., HTC VIVE Pro, was employed as the display device, which has a screen with an original resolution of 2 880×1 600 pixels, a refresh rate of 90 Hz, and a horizontal FOV of 110 degrees. Moreover, a 360-degree video player with the Equirectangular projection was developed to display the videos on the HMD. The display FOV, length of the MTP latency, and audio latency can be set as desired. Our study adopted a single-stimuli scoring strategy[38].

3.2.2 Experiment 1: Obtaining Perceptual Video Quality

In this experiment, ten YUV420 original videos were employed to form a video database, including four 360-degree videos (i.e., denoted as O1 to O4) proposed by Joint Video Exploration Team (JVET) of ITU-T VCEG and ISO/IEC MPEG[39–40] and six 2D videos (i.e., denoted as V1 to V6) provided by the Ultra Video Group[41], as shown in Fig. 2. The 360-degree videos have a spatial resolution of 3 840×1 920 pixels, a framerate of 30 fps and a length of 10 s. The 2D videos have a spatial resolution of 3 840×2 160 pixels, a framerate of 120 fps and a length of 5 s. The experiment was divided into 2 sub-experiments, which were designed to investigate the impact of bitrate and frame rate on the perceptual video quality. Details of the experiment settings are introduced as follows:

1) Investigating the impact of video bitrate

Four 360-degree videos, i.e., O1 to O4, were utilized to investigate the relationship between the video bitrate and the perceptual video quality. The bits per pixel (BPP) were employed to unify the coding bitrates under different resolutions. It can be calculated by

$$\mathrm{BPP} = \frac{Br}{R_H \times R_V \times f}, \tag{1}$$

where Br and f are the video bitrate and frame rate, respectively. R_H and R_V are the horizontal and vertical source resolutions. The original 360-degree videos were down-sampled and encoded using an x265 encoder according to the settings listed in Table 1.

During the experiment, video sequences were displayed in random orders using the HMD. Subjects can change their viewport by rotating their head. There was a 10-second interval between each two video sequences. Subjects could rate the perceptual video quality using the Absolute Category Rating (ACR) 5-point scale (corresponding to the perceived quality of "excellent," "good," "fair," "poor," and "bad" from 5 to 1 point) during the 10-second interval. Before the formal test, the subjects were asked to rate a few example videos to get familiar with the scoring scale and the scoring tool.

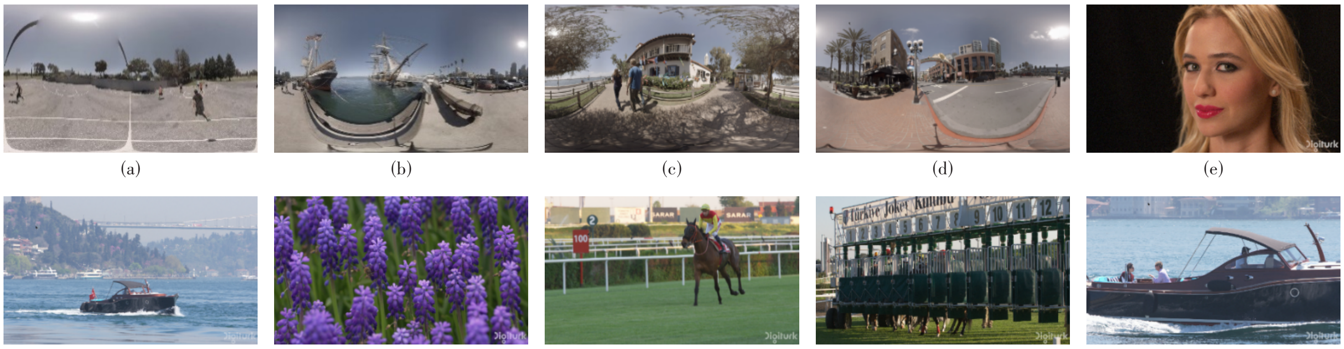
2) Investigating the impact of frame rate

To the best of our knowledge, there is no 360-degree video database containing videos with a frame rate higher than 60 fps. As the screen refresh rate of the current HMDs can reach 90 Hz, we have to use six 2D videos with a high frame rate, i.e., V1 to V6, to study the impact of frame rate. Each original video was repeated twice to generate a video of 10 s. Then, they were down-sampled to 60 fps, 30 fps, and 15 fps. These videos (including the original 120 fps) were further spatially down-sampled to 960 × 540. Videos generated from V1 to V4 were encoded with a fixed quantization parameter (QP), i.e., 22, us-

▼Table 1. Experimental setup

BPP	Bitrate/(Mbit/s)			BPP	Bitrate/(Mbit/s)
	720P (1 280×640)	1080P (1 920×960)	2K (2 160×1 080)		4K (3 840×1 920)
0.016	0.39	0.89	1.12	0.011	2.50
0.032	0.79	1.77	2.24	0.024	5.20
0.056	1.38	3.10	3.92	0.056	12.39
0.08	1.97	4.42	5.60	0.08	17.70
0.16	3.93	8.85	11.20	0.16	35.39
0.20	4.92	11.06	14.00	0.20	44.24

BPP: bits per pixel


(a) (b) (c) (d) (e)
(f) (g) (h) (i) (j)

▲Figure 2. Content of test sequences: (a) Basketball, (b) Harbor, (c) KiteFlite, (d) Gaslamp, (e) Beauty, (f) Bosphorus, (g) Honeybee, (h) Jockey, (i) ReadySetGo, and (j) YachRide ing the x.265 encoder to generate high-quality videos. To investigate whether the QP can influence the impact of framerates on the perceptual video quality, V5 and V6 were encoded with four different QPs, i.e., 22, 32, 36, and 39, to generate four quality levels. During the experiment, video sequences were displayed in their resolution in random orders. It is noted that videos with 120 fps were displayed at 90 fps on the HMD since the refresh rate of the HMD is only 90 Hz.

3.2.3 Experiment 2: Obtaining Visual Realism

Three high-quality stereoscopic videos (3 840×3 840 resolution) were downloaded. Note that the audio tracks were not used in this experiment. These videos last for 20 s and have a frame rate of 30 fps. The projection mode is equirectangular. They were firstly separated into two monoscopic videos, namely the left and right videos, separately. To investigate the impact of stereoscopic vision, the left videos and stereoscopic videos were utilized as the test materials that were further encoded into three quality levels: 1 Mbit/s, 5 Mbit/s and 14 Mbit/s for monoscopic videos and 2 Mbit/s, 8 Mbit/s and 18 Mbit/s for stereoscopic videos. The FOV was set to be 60 degrees, 90 degrees and 110 degrees, respectively. The ACR 5-point scale was also used in this experiment to record the evaluation scores for the perceptual video quality and visual realism. To obtain visual realism, the subjects were asked a question: "To what extent are your visual experiences in the virtual environment consistent with that in the real world?".

3.2.4 Experiment 3: Obtaining Perceptual Audio Quality

The audio tracks from the perceptual evaluation of audio quality (PEAQ) conformance test listed in ITU-R BS.1387[42] were employed as the reference. More specifically, six samples, four music pieces and two speeches, were used, as summarized in Table 2. The sampling frequency of all audio files is 48 kHz. Stereo (two-channel) audio files were used for the test. They were encoded using the Advanced Audio Codec (AAC) encoder with a bit rate of 8 kbit/s, 16 kbit/s, 32 kbit/s, 64 kbit/s, 128 kbit/s, 256 kbit/s, and 320 kbit/s, respectively and a sampling rate of 48 kHz. The generated audio sequences were displayed to subjects on a high-fidelity headphone in a random order. After each display, the subjects were asked to rate the quality levels of audio files in ACR 5-point scales.

3.2.5 Experiment 4: Obtaining Acoustic Realism

The left videos in Experiment 2 encoded with 14 Mbit/s and corresponding audio files were used in this experiment to investigate the influence of the audio quality and spatial audio on acoustic realism. The audio component of these videos was in eight channels with each representing the sound from one direction. The original audio files were encoded using the AAC codec with a bit rate of 128 kbit/s and a sampling rate of 44.1 kHz. The sound from front-left and front-right was firstly mixed into the stereo audio. Then, the stereo audio files and original spatial audio files were encoded with 16 kbit/s, 32 kbit/s, 64 kbit/s, and 128 kbit/s to generate four quality levels. After the display of each audiovisual sequence, two questions were asked: "How do you rate the quality of the audio you just heard?" and "To what extent are your acoustic experiences in the virtual environment consistent with that in the real world?". Then, the subjects used the ACR 5-point scale to score the audio quality and acoustic realism of the test sequences separately.

▼Table 2. Experimental setup

File Name	Signal Type	File Name	Signal Type
FCODSB1.WAV	music	LCODPIP.WAV	music
GCODCLA.WAV	music	NCODSFE.WAV	speech
LCODHRP.WAV	music	KREFSME.WAV	speech

3.2.6 Experiment 5: Obtaining Proprioceptive Matching

In this experiment, the influence of the MTP latency and AL on proprioceptive matching was investigated. First, three left vision videos in Experiment 2 with "excellent" video quality were displayed with seven lengths of MTP latency, i.e., 0 ms, 20 ms, 60 ms, 100 ms, 200 ms, 300 ms, and 500 ms, in a random order. Their audio files (high quality, 128 kbit/s) were displayed with no audio latency. Then, these videos were displayed with no MTP latency while the corresponding spatial audio files (high quality, 128 kbit/s) were displayed with eight different lengths of audio latency, i.e., 0 ms, 20 ms, 60 ms, 150 ms, 300 ms, 500 ms, 1 000 ms, and 2 000 ms, respectively. The subjects were asked to score the degree of proprioceptive matching for the test sequences with the ACR 5-point scale.

3.2.7 Experiment 6: Obtaining Spatial Presence

As listed in Table 3, the original stereoscopic videos (i.e., denoted as S1 to S3) and corresponding stereo audio files in Experiment 2 were first encoded and displayed on the HMD with no MTP latency and AL. Then, the original audiovisual files were encoded with high quality and displayed with six MTP latencies, i.e., 0 ms, 20 ms, 80 ms, 150 ms, 300 ms, and

▼Table 3. Experimental setup

No.	Video / (Mbit/s)	Audio / (kbit/s)	No.	Video / (Mbit/s)	Audio / (kbit/s)
S1	2	16	S2	8	16
S1	8	32	S2	18	32
S1	18	64	S2	2	64
S1	18	16	S3	2	32
S1	2	64	S3	8	64
S1	4	128	S3	18	128
S2	2	16	S3	8	16
S2	8	64	S3	18	32
S2	18	128	S3	2	128

500 ms, respectively. We adopted the 5-point spatial presence scale proposed in Ref. [43] where a point from 5 to 1 indicates the degree of being there from "very strong" to "not at all". The question designed in the experiment was "To what extent did you feel like you were really inside the virtual environment?".

After the subjective tests, the reliability of the subjective results in each experiment was checked using the Pearson Linear Correlation Coefficient (PLCC) adopted by ITU-T Recommendation P.913[38]. According to the suggested threshold of 0.75[38], only the results from two subjects were discarded.

4 Spatial Presence Assessment Model

In the previous section, we construct several test scenarios under different impact factor settings and launched subjective experiments to obtain users' rating scores. These scores are the ground truth of spatial presence under different impact factor settings. In this section, the characteristic of users' perception in each module is analyzed based on the preliminary observation of the experiment results. The weight of each impact factor is determined using the linear regression method.

4.1 Perceptual Video Quality Assessment Module

As studied in Refs. [44] and [45], the impact of frame rate and quantization is separable. We follow this conclusion and hypothesize that the perceptual video quality can be predicted as follows:

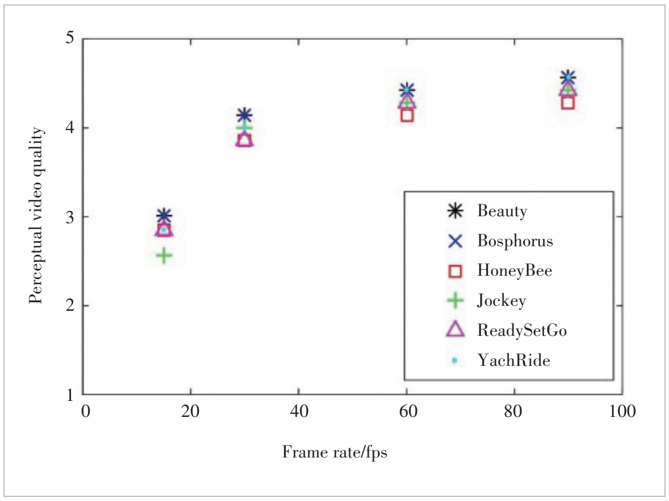
$$\mathrm{PVQ}\left(\mathrm{BPP},f\right)=\mathrm{SQF}\left(\mathrm{BPP}\right)\cdot\mathrm{TCF}\left(f\right), \tag{2}$$

where f represents the frame rate and BPP is the bits per pixel. SQF and TCF are the spatial quality factor and temporal correction factor, respectively. The first term SQF (BPP) measures the quality of encoded frames without considering the impact of frame rate. The second term models how the Mean Opinion Score (MOS) varies with the change of frame rate.

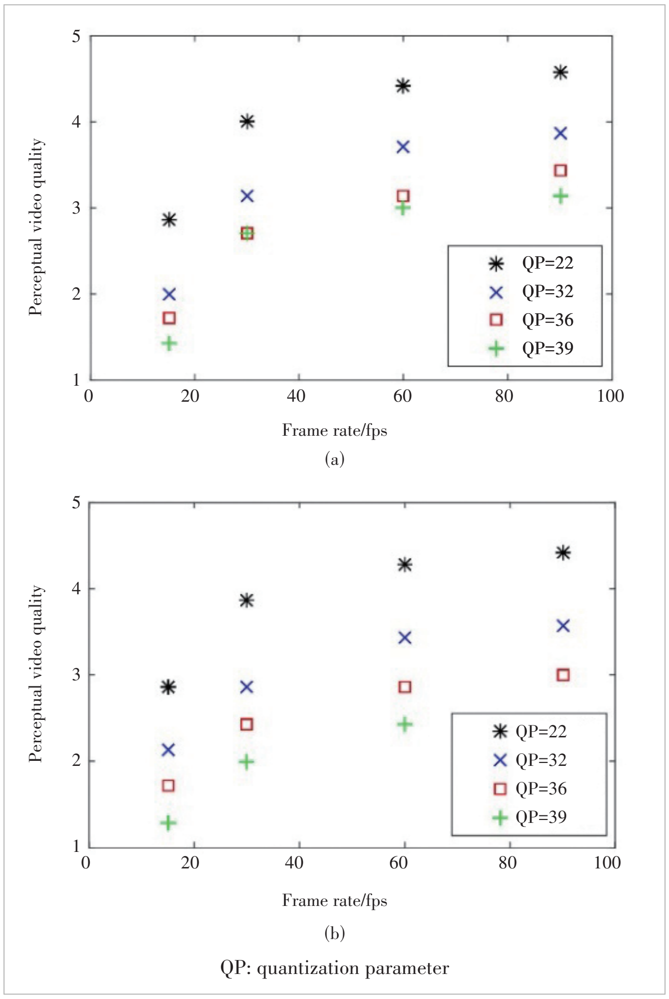
4.1.1 Temporal Correction Factor

Fig. 3 shows the relationship between the frame rate and the perceptual video quality. We can see that the perceptual video quality increases along with the rise of frame rate. Fig. 4 presents the experimental results of the two videos encoded with four different QPs. It can be found that no matter what the QP level is, MOS reduces consistently as the frame rate decreases. In order to examine whether the decreasing trend of MOS against the frame rate is independent of the QP, the MOS scores were normalized and shown in Fig. 5, where the normalized MOS (NMOS) is the ratio of the MOS with the MOS at 30 fps. More specifically, the NMOS is calculated as

$$\mathrm{NMOS}\left(\mathrm{QP},f\right)=\frac{\mathrm{MOS}\left(\mathrm{QP},f\right)}{\mathrm{MOS}\left(\mathrm{QP},30\right)}. \tag{3}$$



▲ **Figure 3. Relationship between the frame rate and perceptual video quality**



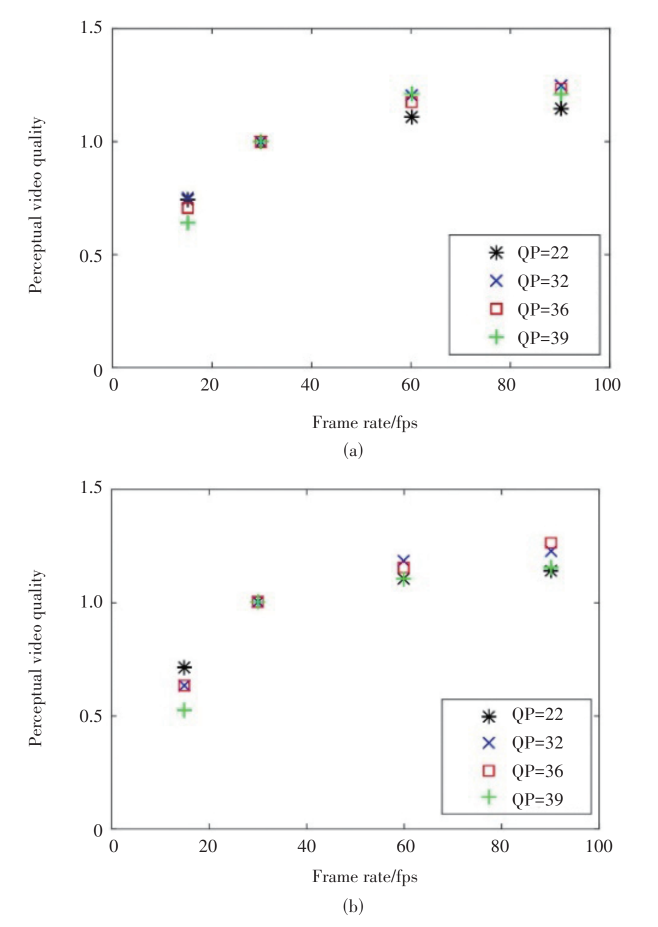
(a)

(b)

QP: quantization parameter

▲ **Figure 4. Experimental results of (a) ReadySetGo and (b) YachRide encoded with four different QPs**

As can be seen in Fig. 5, these NMOS scores corresponding to different QPs almost overlap with each other, indicating



QP: quantization parameter

▲ **Figure 5. Relationship between the frame rate and normalized Mean Opinion Score (NMOS): (a) ReadySetGo and (b) YachRide**

that the decrease of MOS with the frame rate is independent of the QP. This observation follows the conclusions drawn in Refs. [44] and [45] and confirms our hypothesis. The trend in Fig. 5 can be fitted using the function as

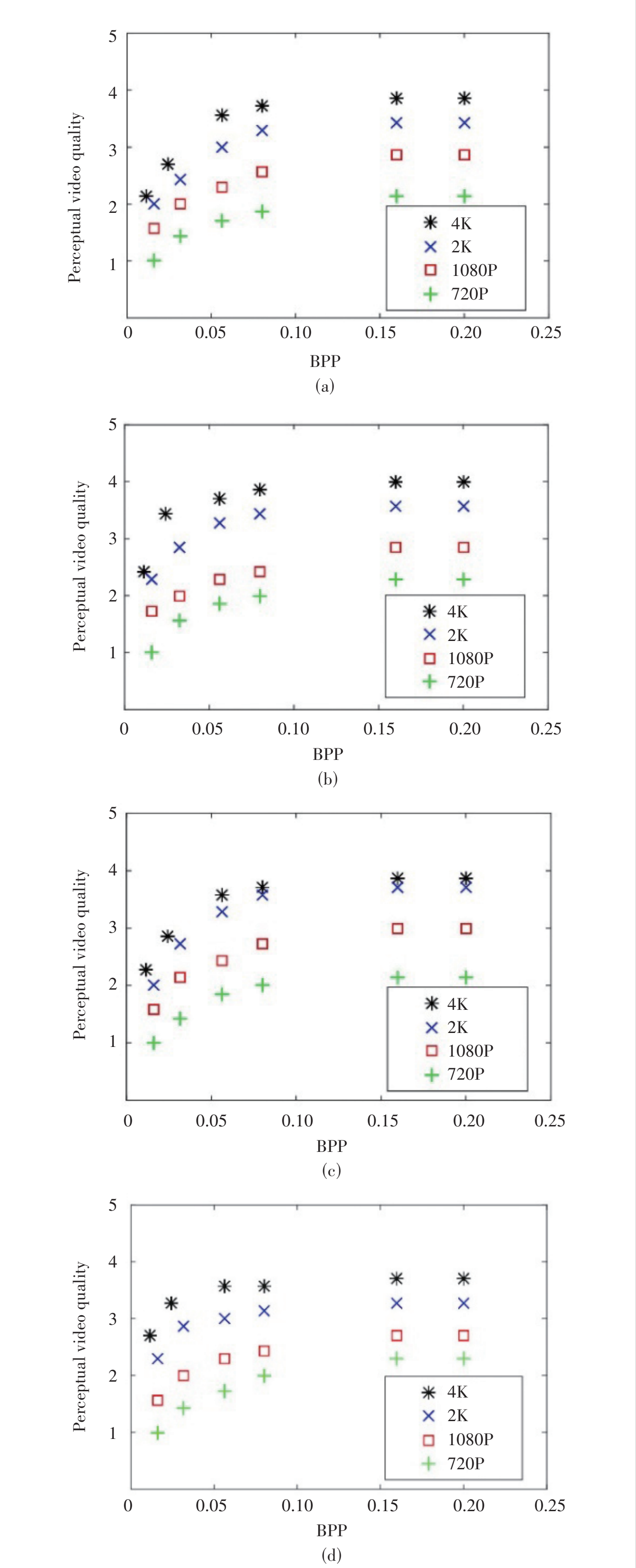
$$\mathrm{TCF}\left(Fr\right) = v_1 \cdot \exp\left(v_2 \cdot f\right) + v_3 \,, \tag{4}$$

where v_1, v_2, and v_3 are −1.672, −0.095 31 and 1.112, respectively, which were obtained by regression.

4.1.2 Spatial Quality Factor

In this subsection, we investigate and modeled the spatial quality, which is mainly influenced by the bitrate, video resolution, and screen resolution. Fig. 6 shows the relationship between the BPP and the perceptual video quality of four 360-degree videos. It can be seen that the perceptual video quality increases with the rise of BPP. However, as for videos at different resolutions, the increasing trends of the perceptual video quality are different. This trend can be represented as

$$\mathrm{SQF}\left(\mathrm{BPP}\right) = v_4 \cdot \ln\left(v_5 \cdot \mathrm{BPP} \cdot 1\,000 + 1\right), \tag{5}$$



BPP: bits per pixel

▲ **Figure 6. Relationship between BPP and perceptual video quality: (a) Basketball, (b) Harbor, (c) KiteFlite, and (d) Gaslamp** where v_4 and v_5 are the model coefficients that can be obtained by regression. The values of v_4 and v_5 are listed in Table 4. It can be seen that the values of v_4 are very close to each other while that of v_5 are quite distinct for different video resolutions. Hence, the average value of v_4 is used as a fixed coefficient. The value of v_5 is then regressed again.

To reflect the impact of video resolution and screen resolution on perceived video quality, we employ the integrated assessment parameter that we proposed in the previous work[46], i.e., the number of effective video pixels per degree (ED-PPD) displayed on the screen of HMD. The effective pixels do not include the pixels interpolated by the up-sampling process. This parameter is calculated as

$$\mathrm{ED-PPD} = \begin{cases} \dfrac{R_H}{360}, & R_H \leqslant R_{SH} \cdot \dfrac{360}{\mathrm{FOV}} \\ \dfrac{R_{SH}}{\mathrm{FOV}}, & R_H > R_{SH} \cdot \dfrac{360}{\mathrm{FOV}} \end{cases}, \tag{6}$$

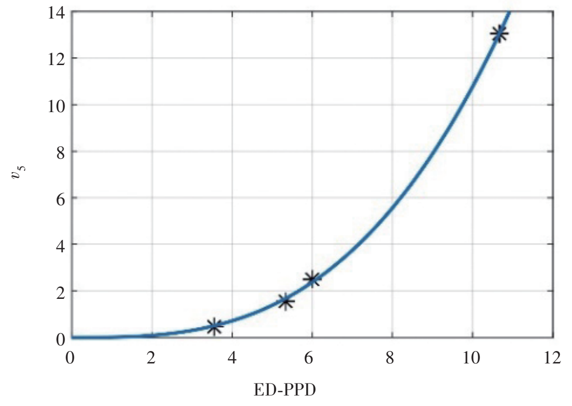
where R_H and R_{SH} are the horizontal resolution of 360-degree video and screen, respectively. When the horizontal pixels of the video displayed on the screen are more than the horizontal pixels on the screen, the ED-PPD will be saturated.

Fig. 7 shows the relationship between the ED-PPD and v_5. It can be seen that the values of v_5 and ED-PPD are in accordance with the power function relationship, which can be expressed as

$$v_5 = v_6 \cdot \mathrm{ED-PPD}^{v_7}, \tag{7}$$

▼**Table 4. Values of v_4 and v_5**

Video Resolution	v_4	v_5
720P	0.497 4	0.525 7
1080P	0.529 2	1.369 0
2K	0.537 1	4.584 0
4K	0.497 4	16.580 0



ED-PPD: effective video pixels per degree

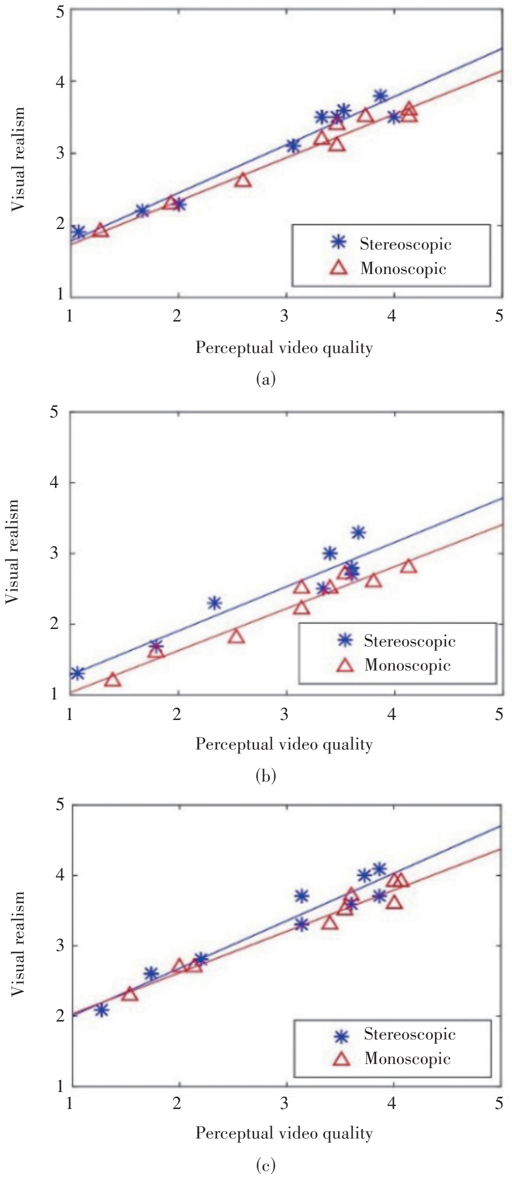
▲**Figure 7. Relationship between ED-PPD and v_5**

where v_6 and v_7 are equal to 0.011 7 and 2.962, respectively.

By substituting Eqs. (4), (5) and (7) into Eq. (2), the perceptual video quality of 360-degree videos can be modeled.

4.2 Visual Realism Assessment

According to the results of Experiment 2, Fig. 8 shows the relationship between perceptual video quality and visual realism. It can be seen that there is a strong correlation between the perceptual video quality and visual realism. For the influence of FOV, it can be observed that a higher FOV leads to a higher visual realism. The Kruskal-Wallis H test showed that there is a significant effect of FOV on visual realism, with p = 0.001 for monoscopic videos and p = 0.039 for stereoscopic



(a)

(b)

(c)

▲**Figure 8. Relationship between the perceptual video quality and visual realism: (a) 60 FOV (field of view), (b) 90 FOV, and (c) 110 FOV** videos. A one-way analysis of variance (ANOVA) test indicates that there is no significant effect of the type of vision on visual realism. Based on the results above, the video quality and FOV appear to have a more significant impact on visual realism than the type of vision. Thus, the relationship of perceptual video quality, FOV, and visual realism can be calculated by

$$\mathrm{VRE}\left(\mathrm{PVQ},\mathrm{FOV}\right)=\max\left(\min\left(v_8\mathrm{PVQ}+v_9\mathrm{FoV}+v_{10},5\right),1\right), \tag{8}$$

where v_8, v_9 and v_{10} are equal to 0.595, 0.02 and −0.735, respectively.

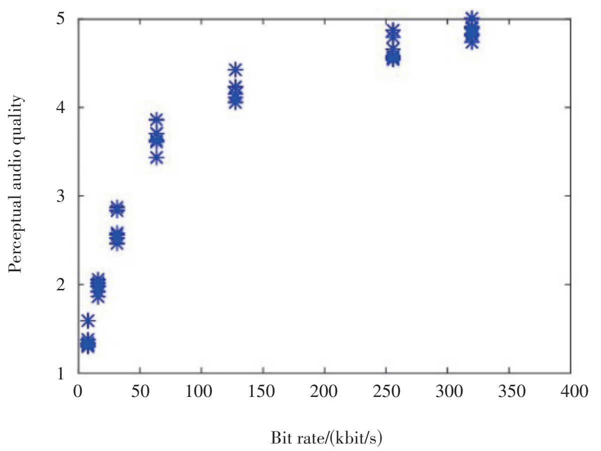
4.3 Perceptual Audio Quality and Acoustic Realism Assessment

We first model the perceptual audio quality using the experimental results of Experiment 3. Fig. 9 shows the logarithmic relationship between the audio bitrate and the perceptual audio quality. This relationship can be represented as

$$\mathrm{PAQ}\left(\mathrm{ABr}\right)=1+v_{11}-\frac{v_{11}}{1+\left(\frac{\mathrm{ABr}}{v_{12}}\right)^{v_{13}}}, \tag{9}$$

where v_{11}, v_{12} and v_{13} are equal to 4.103, 42.36 and 1.251, respectively.

As for AR, Fig. 10 shows the relationship between the perceptual audio quality and acoustic realism. It can be found that there is a significant linear relationship between the audio quality and acoustic realism for stereo audio (R^2 = 0.881, F = 213.251, and p = 0.000 < 0.05) and for spatial audio (R^2 = 0.955, F = 73.791, and p = 0.000 < 0.05). The relationship in Fig.10 can be expressed as



▲ **Figure 9. Relationships between the audio bit rate and perceptual audio quality**

$$\mathrm{AR}\left(\mathrm{PAQ}\right)=v_{14}\mathrm{PAQ}+v_{15}, \tag{10}$$

where v_{14} and v_{15} are equal to 0.733 and 0.634 for the stereo audio, and equal to 0.682 and 1.167 for the spatial audio.

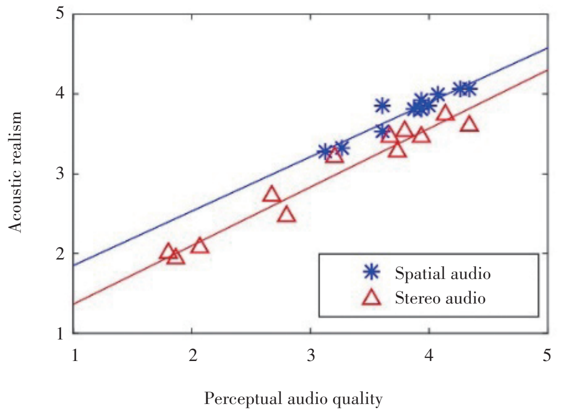
4.4 Proprioceptive Matching Assessment

Fig. 11 shows the relationship between the two types of delay and proprioceptive matching. It can be seen that the proprioceptive matching decreases with the increase of both the MTP latency and AL. Here, the degradations of proprioceptive matching caused by the MTP latency and AL are calculated by

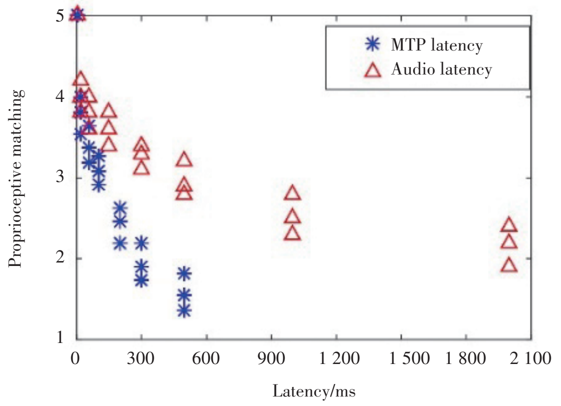
$$\mathrm{DMOS}\left(\mathrm{MTP}\right)=5-\mathrm{MOS}\left(\mathrm{MTP}\right), \tag{11}$$

$$\mathrm{DMOS}\left(\mathrm{AL}\right)=5-\mathrm{MOS}\left(\mathrm{AL}\right). \tag{12}$$

Fig. 12 shows the relationship between the two types of delay and the degradation of proprioceptive matching. This rela-

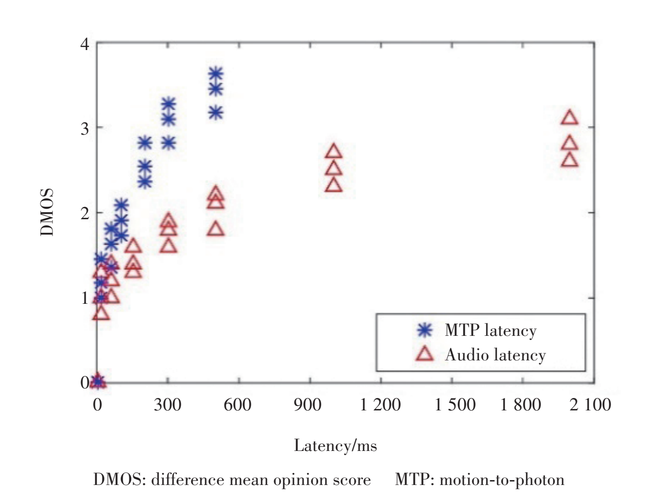


▲ **Figure 10. Relationship between the perceptual audio quality and acoustic realism rated on Head-Mounted Display (HMD)**



MTP: motion-to-photon

▲ **Figure 11. Relationship between the two types of delay and the proprioceptive matching**



▲**Figure 12. Relationship between the two types of latency and the degradation of proprioceptive matching**

tionship can be represented by

$$\mathrm{DMOS}(\mathrm{MTP}) = \max\left(\min\left(\ln\left(v_{16}\mathrm{MTP} + 1\right), 4\right), 0\right), \tag{13}$$

$$\mathrm{DMOS}(\mathrm{AL}) = \max\left(\min\left(v_{17}\ln\left(v_{18}\mathrm{AL} + 1\right), 4\right), 0\right), \tag{14}$$

where v_{16}, v_{17} and v_{18} are equal to 0.065 46, 0.428 9 and 0.275 4, respectively. We modeled the proprioceptive matching as

$$\mathrm{PM}(\mathrm{MTP,AL}) = \max\left(\min\left(5 - \mathrm{DMOS}(\mathrm{MTP}) - \mathrm{DMOS}(\mathrm{AL}), 5\right), 1\right). \tag{15}$$

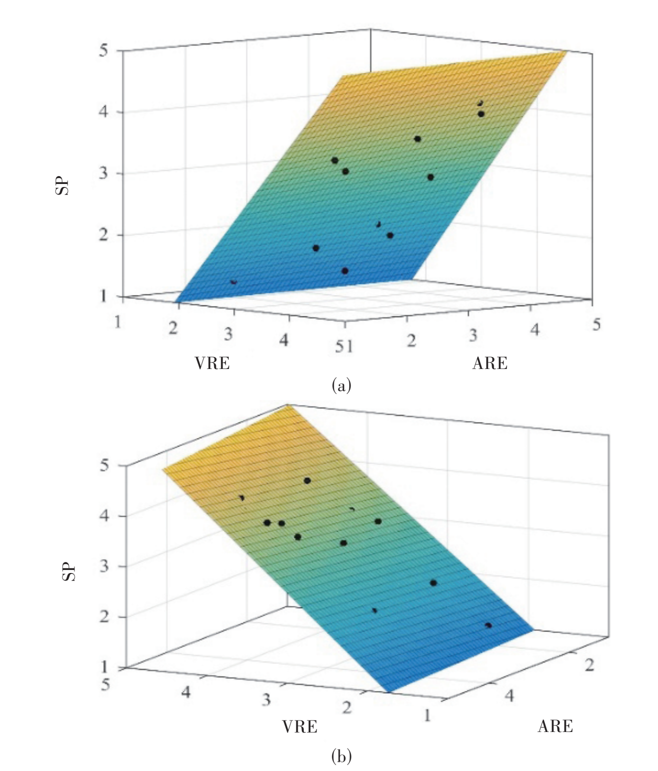
4.5 Spatial Presence Assessment

First, the relationship between the visual/acoustic realism and the spatial presence is modeled. As shown in Fig. 13, the spatial presence increases with the rise of VRE and ARE. This phenomenon confirms the conclusion drawn in our previous work[18]. The relationship shown in Fig. 13 can be calculated as

$$\mathrm{SPAV}(\mathrm{VRE}, \mathrm{ARE}) = \min\left(\max\left(v_{19}\mathrm{VRE} + \mathrm{v}_{20}\mathrm{ARE} + v_{21}\mathrm{VRE} \times \mathrm{ARE} + v_{22}, 1\right), 5\right), \tag{16}$$

where v_{19}, v_{20}, v_{21}, and v_{22} are equal to 1.285, 0.01, 0.027 4, and −1.529, respectively; SPAV represents the spatial presence provided by the visual and acoustic experience.

Second, the impact of proprioceptive matching is investigated. Fig. 14 shows the relationship between the proprioceptive matching and the degradation of spatial presence. We can find that the degradation of spatial presence decreases with the increase of proprioceptive matching. The relationship in Fig. 14 can be modeled as

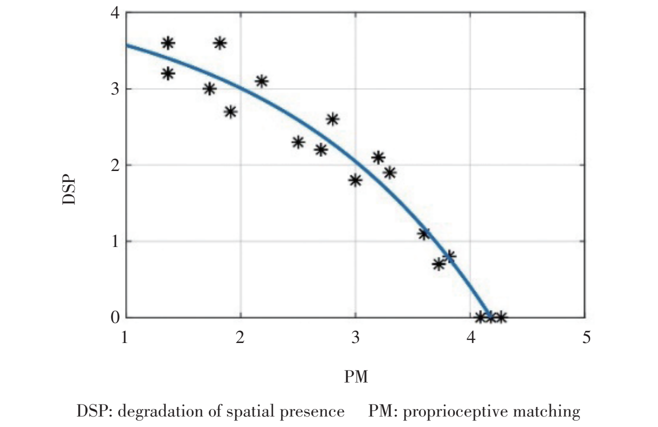


(a)

(b)

ARE: acoustic realism SP: spatial presence VRE: visual realism

▲**Figure 13. Relationships between the two types of realism and the spatial presence**



▲ **Figure 14. Relationships between the proprioceptive matching and degradation of spatial presence**

$$\mathrm{DSP}(\mathrm{PM}) = v_{23} \cdot \exp\left(v_{24} \cdot \mathrm{PM}\right) + v_{25}, \tag{17}$$

where v_{23}, v_{24} and v_{25} are equal to −0.467 9, 0.533 8 and 4.367, respectively. Hence, the spatial presence can be calculated by

$$\mathrm{SP}(\mathrm{SPAV}, \mathrm{DSP}) = \min\left(\max\left(\mathrm{SPAV} - \mathrm{DSP}, 1\right), 5\right). \tag{18}$$

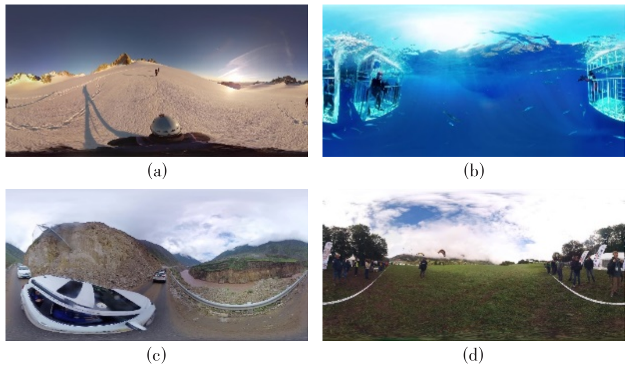
By utilizing the proposed model, the spatial presence of

360-degree video can be assessed based on the corresponding technical parameters extracted from the VR system.

5 Performance Evaluation

The performance of the proposed model was evaluated on a test set consisting of another four YUV420 360-degree video sequences that had a video resolution of 3 840×1 920 and a video framerate of 30 fps. Screenshots of the video content are shown in Fig. 15. Four lossless audio files (PCM, 48 kHz) containing two channels were utilized as the background sound of these 360-degree videos. The 360-degree videos were firstly down-sampled to 2K resolution and encoded with a BPP of 0.02, 0.06, 0.14, and 0.19 using the x.265 encoder. The audio files were encoded with 16 kbit/s, 64 kbit/s, 128 kbit/s, and 256 kbit/s using the AAC codec. We conducted two experiments to verify the performance of the proposed model by changing the video bitrate, audio bitrate, and MTP latency. In the first experiment, audiovisual files were displayed without MTP latency. The display FOV was set to be 90 degrees and 110 degrees, respectively. The details of the setting are shown in Table 5. In the second experiment, audiovisual files with 4K resolution were displayed with three MTP latencies, i. e., 40 ms, 120 ms, and 260 ms, respectively. The display FOV was set to be 110 degrees. The details of the setting are shown in Table 6. A total number of 30 subjects participated in these two experiments. After each display, the subjects provided their ratings on the spatial presence on a five-point scale.

Since there is no model evaluating the spatial presence that can be used as a comparison, we only show the performance of the proposed model. The performance is evaluated in two ways: 1) comparing predicted scores of the spatial presence with the subjective MOS, and 2) comparing the predicted scores with the subjective scores rated by individuals.



▲Figure 15. Content of test sequences: (a) Driving, (b) Shark, (c) Glacier, and (d) Paramotor

▼Table 5. Setup for the video

Video (BPP)	Audio/(kbit/s)
0.02	16, 64, 128, 256
0.06	16, 64, 128, 256
0.14	16, 64, 128, 256
0.19	16, 64, 128, 256

BPP: bits per pixel

▼Table 6. Setup for the audio

Video (BPP)	Audio/(kbit/s)
0.06	16, 64, 256
0.14	16, 64, 256
0.19	16, 64, 256

BPP: bits per pixel

5.1 Predicted Scores vs MOS

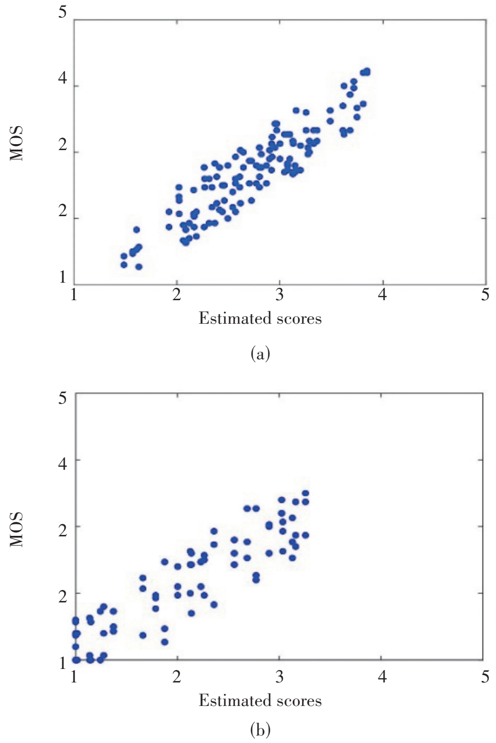
Three commonly used performance criteria are employed to measure the performance of the proposed model: Pearson Correlation Coefficient (PCC), Root-Mean-Squared Error (RMSE), and Spearman Rank Order Correlation Coefficient (SROCC).

The model performance is given in Table 7. It can be found that reliable prediction performance is obtained when using the proposed spatial presence evaluation model.

▼Table 7. Experimental results

Experiment	PCC	SROCC	RMSE
1	0.910	0.894	0.277
2	0.908	0.900	0.335

PCC: Pearson Correlation Coefficient
RMSE: Root-Mean-Squared Error
SROCC: Spearman Rank Order Correlation Coefficient



MOS: Mean Opinion Score

▲Figure 16. Scatter plots of the subjective spatial presence versus predicted objective scores: (a) result of Experiment 1 and (b) result of Experiment 2

To visualize the performance, Fig. 16 shows the scatter plots of objective scores predicted by the proposed model against the subjective MOSs. This figure clearly shows that the proposed model exhibits good convergence and monotonicity performance.

5.2 Predicted Scores vs Individual Ratings

To also check the accuracy of the proposed model, we evaluated the performance of the model against the individual ratings of subjects. Again, PCC, SROCC, and RMSE were calculated. For Experiment 1, we found that the PCC, SROCC, and RMSE ranged from 0.882 to 0.926, 0.878 to 0.922, and 0.443 to 0.227, respectively. For Experiment 2, we found that the PCC, SROCC, and RMSE ranged between 0.886 to 0.924, 0.881 to 0.918, and 0.462 to 0.214. Among the 30 subjects, the lowest, medium and highest prediction results are shown in Table 8. It can be found that a relatively good prediction performance is always guaranteed using the proposed model.

We also calculated the percentage that the predicted scores match the subjective scores to better verify the accuracy of the proposed model. A match is found if a predicted score (after the rounding process) is the same as the subjective score rated by the participants. The results show that the proposed model matches the subjective ratings with an accuracy of 83.7% and 82.4% for Experiments 1 and 2, respectively. It can be concluded that the proposed model manifests itself as a reliable spatial presence indicator that can be directly used in current 360-degree video applications.

▼Table 8. Model performance

Experiment	Subject No. 1			Subject No. 2			Subject No. 3		
	PCC	SROCC	RMSE	PCC	SROCC	RMSE	PCC	SROCC	RMSE
1	0.882	0.878	0.443	0.926	0.922	0.227	0.908	0.902	0.282
2	0.886	0.881	0.462	0.924	0.918	0.214	0.904	0.898	0.343

PCC: Pearson Correlation Coefficient
RMSE: Root-Mean-Squared Error
SROCC: Spearman Rank Order Correlation Coefficient

6 Conclusions

In this paper, we propose a spatial presence assessment framework for measuring users' sense of spatial presence in 360-degree video services. Well-designed subjective experiments are conducted to obtain accurate subjective ratings of spatial presence. An objective spatial presence prediction model is further proposed. Experimental results show that the proposed model can achieve good prediction accuracy in terms of PCC, SROCC, and RMSE. The proposed scheme serves as guidelines for the research community to better understand the spatial presence perception. It also provides valuable recommendations for the industry to further improve its quality of service.

References

[1] DELOITTE. Digital democracy survey: a multi-generational view of consumer technology, media and telecom trends, digital democracy survey 9th edition [R]. 2022

[2] ZHU W H, ZHAI G T, TAO M X, et al. Quality of experience estimation of ultra-high definition content [J]. ZTE technology journal, 2021, 27(1): 37 – 43. DOI: 10.12142/ZTETJ.202101009

[3] LI J L, ZHAO X, YANG Y. A review of interactive video quality assessment methods [J]. ZTE technology journal, 2021, 27(1): 44 – 47. DOI: 10.12142/ZTETJ.202101010

[4] LOMBARD M, JONES M T. Defining presence [M]//Immersed in media. Cham: Springer International Publishing, 2015: 13 – 34. DOI: 10.1007/978-3-319-10190-3_2

[5] SCHUEMIE M J, VAN DER STRAATEN P, KRIJN M, et al. Research on presence in virtual reality: a survey [J]. CyberPsychology & behavior, 2001, 4(2): 183 – 201. DOI: 10.1089/109493101300117884

[6] SEO Y, KIM M, JUNG Y, et al. Avatar face recognition and self-presence [J]. Computers in human behavior, 2017, 69: 120 – 127. DOI: 10.1016/j.chb.2016.12.020

[7] FELTON W M, JACKSON R E. Presence: a review [J]. International journal of human-computer interaction, 2022, 38(1): 1 – 18. DOI: 10.1080/10447318.2021.1921368

[8] LOMBARD M, DITTON T. At the heart of it all: the concept of presence [J]. Journal of computer-mediated communication, 2006, 3(2). DOI: 10.1111/j.1083-6101.1997.tb00072.x

[9] NORTH M M, NORTH S M. A comparative study of sense of presence of virtual reality and immersive environments [J]. Australasian journal of information systems, 2016, 20. DOI: 10.3127/ajis.v20i0.1168

[10] GONÇALVES G, MELO M, BARBOSA L, et al. Evaluation of the impact of different levels of self-representation and body tracking on the sense of presence and embodiment in immersive VR [J]. Virtual reality, 2022, 26(1): 1 – 14. DOI: 10.1007/s10055-021-00530-5

[11] SKARBEZ R, BROOKS Jr F P, WHITTON M C. A survey of presence and related concepts [J]. ACM computing surveys (CSUR), 2017, 50(6): 1 – 39. DOI: 10.1145/3134301

[12] LAARNI J, RAVAJA N, SAARI T, et al. Ways to measure spatial presence: review and future directions [M]//Immersed in media. Cham: Springer International Publishing, 2015: 139 – 185. DOI: 10.1007/978-3-319-10190-3_8

[13] AL-JUNDI H A, TANBOUR E Y. A framework for fidelity evaluation of immersive virtual reality systems [J].Virtual reality, 2022, 26(3): 1103 – 1122. DOI: 10.1007/s10055-021-00618-y

[14] EGAN D, BRENNAN S, BARRETT J, et al. An evaluation of heart rate and electrodermal activity as an objective QoE evaluation method for immersive virtual reality environments [C]//Proc. 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX). IEEE, 2016: 1 – 6. DOI: 10.1109/QoMEX.2016.7498964

[15] CHESSA M, MAIELLO G, BORSARI A, et al. The perceptual quality of the oculus rift for immersive virtual reality [J]. Human-computer interaction, 2019, 34(1): 51 – 82. DOI: 10.1080/07370024.2016.1243478

[16] TERKILDSEN T, MAKRANSKY G. Measuring presence in video games: an investigation of the potential use of physiological measures as indicators of presence [J]. International journal of human-computer studies, 2019, 126: 64 – 80. DOI: 10.1016/j.ijhcs.2019.02.006

[17] GRASSINI S, LAUMANN K. Questionnaire measures and physiological correlates of presence: a systematic review [J]. Frontiers in psychology, 2020, 11: 349. DOI: 10.3389/fpsyg.2020.00349

[18] ZOU W J, YANG F Z, ZHANG W, et al. A framework for assessing spatial presence of omnidirectional video on virtual reality device [J]. IEEE access, 2018, 6: 44676 – 44684. DOI: 10.1109/ACCESS.2018.2864872

[19] WITMER B G, SINGER M J. Measuring presence in virtual environments: a presence questionnaire [J]. Presence: teleoperators and virtual environments, 1998, 7(3): 225 – 240. DOI: 10.1162/105474698565686

[20] LESSITER J, FREEMAN J, KEOGH E, et al. A cross-media presence questionnaire: the ITC-sense of presence inventory [J]. Presence: teleoperators and virtual environments, 2001, 10(3): 282 – 297. DOI: 10.1162/105474601300343612

[21] JENNETT C, COX A L, CAIRNS P, et al. Measuring and defining the experience of immersion in games [J]. International journal of human-computer studies, 2008, 66(9): 641 – 661. DOI: 10.1016/j.ijhcs.2008.04.004

[22] VORDERER P, WIRTH W, GOUVEIA F R, et al. MEC spatial presence questionnaire (MEC-SPQ): short documentation and instructions for application [R]. 2004

[23] CUMMINGS J J, WERTZ E E. Capturing social presence: Concept explication through an empirical analysis of social presence measures [J]. Journal of computer-mediated communication, 2022, 28(1): zmac027. DOI: 10.1093/jcmc/zmac027

[24] LIN J J W, DUH H B L, PARKER D E, et al. Effects of field of view on presence, enjoyment, memory, and simulator sickness in a virtual environment [C]//Proc. IEEE Virtual Reality. IEEE, 2002: 164 – 171. DOI: 10.1109/VR.2002.996519

[25] YU M, LAKSHMAN H, GIROD B. A framework to evaluate omnidirectional video coding schemes [C]//Proc. 2015 IEEE International Symposium on Mixed and Augmented Reality. IEEE, 2015: 31 – 36. DOI: 10.1109/ISMAR.2015.12

[26] ZAKHARCHENKO V, CHOI K P, PARK J H. Quality metric for spherical panoramic video [C]//Proc. SPIE Optics and Photonics for Information Processing X. SPIE, 2016. DOI: 10.1117/12.2235885

[27] XU M, LI C, LIU Y F, et al. A subjective visual quality assessment method of panoramic videos [C]//Proc. 2017 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2017: 517 – 522. DOI: 10.1109/ICME.2017.8019351

[28] UPENIK E, ŘEŘÁBEK M, EBRAHIMI T. Testbed for subjective evaluation of omnidirectional visual content [C]//Proc. 2016 Picture Coding Symposium (PCS). IEEE, 2017: 1 – 5. DOI: 10.1109/PCS.2016.7906378

[29] SCHATZ R, SACKL A, TIMMERER C, et al. Towards subjective quality of experience assessment for omnidirectional video streaming [C]//Proc. Ninth International Conference on Quality of Multimedia Experience (QoMEX). IEEE, 2017: 1 – 6. DOI: 10.1109/QoMEX.2017.7965657

[30] OZCINAR C, CABRERA J, SMOLIC A. Visual attention-aware omnidirectional video streaming using optimal tiles for virtual reality [J]. IEEE journal on emerging and selected topics in circuits and systems, 2019, 9(1): 217 – 230. DOI: 10.1109/JETCAS.2019.2895096

[31] GHAZNAVI-YOUVALARI R, ZARE A, AMINLOU A, et al. Shared coded picture technique for tile-based viewport-adaptive streaming of omnidirectional video [J]. IEEE transactions on circuits and systems for video technology, 2019, 29(10): 3106 – 3120. DOI: 10.1109/TCSVT.2018.2874179

[32] DUAN H Y, ZHAI G T, MIN X K, et al. Perceptual quality assessment of omnidirectional images [C]//IEEE International Symposium on Circuits and Systems (ISCAS). 2018: 1 – 5. DOI: 10.1109/ISCAS.2018.8351786

[33] XU M, LI C, CHEN Z Z, et al. Assessing visual quality of omnidirectional videos [J]. IEEE transactions on circuits and systems for video technology, 2019, 29(12): 3516 – 3530. DOI: 10.1109/TCSVT.2018.2886277

[34] ERMI L, MÄYRÄ F. Fundamental components of the gameplay experience: analysing immersion [C]//Digital Games Research Conference 2005, Changing Views: Worlds in Play. DBLP, 2005: 37 – 53

[35] SLATER M, WILBUR S. A framework for immersive virtual environments (FIVE): speculations on the role of presence in virtual environments [J]. Presence: teleoperators and virtual environments, 1997, 6(6): 603 – 616. DOI: 10.1162/pres.1997.6.6.603

[36] CUMMINGS J J, BAILENSON J N. How immersive is enough? A meta-analysis of the effect of immersive technology on user presence [J]. Media psychology, 2016, 19(2): 272 – 309. DOI: 10.1080/15213269.2015.1015740

[37] ZHAO J B, ALLISON R S, VINNIKOV M, et al. Estimating the motion-to-photon latency in head mounted displays [C]//Proc. 2017 IEEE Virtual Reality (VR). IEEE, 2017: 313 – 314. DOI: 10.1109/VR.2017.7892302

[38] ITU. Methods for the subjective assessment of video quality, audio quality and audiovisual quality of internet video and distribution quality television in any environment: ITU-T recommendation P.913 [S]. 2021

[39] ASBUN E, HE Y, HE Y, et al. AHG8: interdigital test sequences for virtual reality video coding [R]. 2016

[40] SUN W, GUO R. Test sequences for virtual reality video coding from letinVR [R]. 2016

[41] MERCAT A, VIITANEN M, VANNE J. UVG dataset: 50/120 fps 4K sequences for video codec analysis and development [C]//Proc. ACM Multimedia Systems Conference. ACM, 2020: 297 – 302. DOI: 10.1145/3339825.3394937

[42] ITU. Methods for objective measurements of perceived audio quality: ITU-R recommendation BS 13871 [S]. 2001

[43] BAILENSON J N, SWINTH K, HOYT C, et al. The independent and interactive effects of embodied-agent appearance and behavior on self-report, cognitive, and behavioral markers of copresence in immersive virtual environments [J]. Presence: teleoperators and virtual environments, 2005, 14(4): 379 – 393. DOI: 10.1162/105474605774785235

[44] OU Y F, MA Z, LIU T, et al. Perceptual quality assessment of video considering both frame rate and quantization artifacts [J]. IEEE transactions on circuits and systems for video technology, 2011, 21(3): 286 – 298. DOI: 10.1109/TCSVT.2010.2087833

[45] OU Y F, LIU T, ZHAO Z, et al. Modeling the impact of frame rate on perceptual quality of video [C]//Proc. 15th IEEE International Conference on Image Processing. IEEE, 2008: 689 – 692. DOI: 10.1109/ICIP.2008.4711848

[46] ZOU W J, YANG F Z, WAN S. Perceptual video quality metric for compression artefacts: from two-dimensional to omnidirectional [J]. IET image processing, 2018, 12(3): 374 – 381. DOI: 10.1049/iet-ipr.2017.0826

Biographies

ZOU Wenjie received his BS and PhD degrees from Xidian University, China in 2009 and 2017, respectively. He is currently a lecturer with the Multimedia Communication Laboratory, Xidian University. His research interests include QoE, video quality assessment, and multimedia compression.

GU Chengming received his BE degree in communication engineering from Xidian University, China in 2021. He is currently working toward an ME degree in information and communication engineering with Xidian University. His research interests include video coding and processing.

FAN Jiawei received his BE degree in telecommunications engineering from Xidian University, China in 2022. He is currently working toward his ME degree in electronic information with Xidian University. His research interests include video coding and processing.

HUANG Cheng (huang.cheng5@zte.com.cn) received his MS degree from the School of Computer Science and Engineering, Southeast University, China. He is currently a senior system architect and project manager of video technology research at ZTE Corporation. His research interests include visual coding, storage, transport, and multimedia systems.

BAI Yaxian received his MS degree in communication engineering from Wuhan University of Technology, China. She is currently a senior engineer at ZTE Corporation. Her research interests include video coding and processing and point cloud compression and transmission.

ZTE Communications

Table of Contents, Volume 21, 2023

Volume-Number-Page

Special Topic

Federated Learning over Wireless Networks

Evolution of AI Enabled Wireless Networks

Volume-Number-Page

Reinforcement Learning and Intelligent Decision

3D Point Cloud Processing and Applications

Research Papers

Review

The 9th Editorial Board of ZTE Communications

The 1st Youth Expert Committee
for Promoting Industry-University-Institute Cooperation

Director **CHEN Wei,** Beijing Jiaotong University

Deputy Director **QIN Xiaoqi,** Beijing University of Posts and Telecommunications

LU Dan, ZTE Corporation

Members (Surname in Alphabetical Order)

CAO Jin	Xidian University
CHEN Li	University of Science and Technology of China
CHEN Qimei	Wuhan University
CHEN Shuyi	Harbin Institute of Technology
CHEN Wei	Beijing Jiaotong University
GUAN Ke	Beijing Jiaotong University
HAN Kaifeng	China Academy of Information and Communications Technology
HE Zi	Nanjing University of Science and Technology
HU Jie	University of Electronic Science and Technology of China
HUANG Chen	Purple Mountain Laboratories
LI Ang	Xi’an Jiaotong University
LIU Chunsen	Fudan University
LIU Fan	Southern University of Science and Technology
LIU Junyu	Xidian University
LU Dan	ZTE Corporation
LU Youyou	Tsinghua University
NING Zhaolong	Chongqing University of Posts and Telecommunications
QI Liang	Shanghai Jiao Tong University
QIN Xiaoqi	Beijing University of Posts and Telecommunications
QIN Zhijin	Tsinghua University
SHI Yinghuan	Nanjing University
WANG Jingjing	Beihang University
WANG Xinggang	Huazhong University of Science and Technology
WANG Yongqiang	Tianjin University
WEN Miaowen	South China University of Technology
WU Yongpeng	Shanghai Jiao Tong University
XIA Wenchao	Nanjing University of Posts and Telecommunications
XU Mengwei	Beijing University of Posts and Telecommunications
XU Tianheng	Shanghai Advanced Research Institute, Chinese Academy of Sciences
YANG Chuanchuan	Peking University
YIN Haifan	Huazhong University of Science and Technology
YU Jihong	Beijing Institute of Technology
ZHANG Jiao	Beijing University of Posts and Telecommunications
ZHANG Yuchao	Beijing University of Posts and Telecommunications
ZHANG Jiayi	Beijing Jiaotong University
ZHAO Yuda	Zhejiang University
ZHOU Yi	Southwest Jiaotong University
ZHU Bingcheng	Southeast University

ZTE COMMUNICATIONS

中兴通讯技术（英文版）

ZTE Communications has been indexed in the following databases:

- Abstract Journal
- Cambridge Scientific Abstracts (CSA)
- China Science and Technology Journal Database
- Chinese Journal Fulltext Databases
- Index of Copurnicus
- Ulrich's Periodicals Directory
- Wanfang Data
- WJCI 2022

ZTE COMMUNICATIONS
Vol. 21 No. 4 (Issue 85, End of Volume)
Quarterly
First English Issue Published in 2003

Supervised by:
Anhui Publishing Group

Sponsored by:
Time Publishing and Media Co., Ltd.
Shenzhen Guangyu Aerospace Industry Co., Ltd.

Published by:
Anhui Science & Technology Publishing House

Edited and Circulated (Home and Abroad) by:
Magazine House of ZTE Communications

Staff Members:
General Editor: WANG Xiyu
Editor-in-Chief: JIANG Xianjun
Executive Editor-in-Chief: HUANG Xinming
Editorial Director: LU Dan
Editor-in-Charge: ZHU Li
Editors: REN Xixi, XU Ye, YANG Guangxi
Producer: XU Ying
Circulation Executive: WANG Pingping
Assistant: WANG Kun

Editorial Correspondence:
Add: 12F Kaixuan Building, 329 Jinzhai Road,
Hefei 230061, P. R. China
Tel: +86-551-65533356
Email: magazine@zte.com.cn
Website: http://zte.magtechjournal.com

Annual Subscription: RMB 120
Printed by:
Hefei Tiancai Color Printing Company
Publication Date: December 25, 2023

China Standard Serial Number: ISSN 1673-5188
CN 34-1294/TN