SST-V: A Scalable Semantic Transmission Framework for Video

LIU Chenyao[1], GUO Jiejie[2], ZHANG Yimeng[1], XU Wenjun[1,3], LIU Yiming[1]

(1. State Key Laboratory of Network and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China;
2. School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China;
3. Department of Mathematics and Theories, Peng Cheng Laboratory, Shenzhen 518066, China)

DOI: 10.12142/ZTECOM.202302010

https://kns.cnki.net/kcms/detail/34.1294.TN.20230524.1802.002.html, published online May 26, 2023

Manuscript received: 2023-02-11

Abstract: The emerging new services in the sixth generation (6G) communication system impose increasingly stringent requirements and challenges on video transmission. Semantic communications are envisioned as a promising solution to these challenges. This paper provides a highly-efficient solution to video transmission by proposing a scalable semantic transmission algorithm, named scalable semantic transmission framework for video (SST-V), which jointly considers the semantic importance and channel conditions. Specifically, a semantic importance evaluation module is designed to extract more informative semantic features according to the estimated importance level, facilitating high-efficiency semantic coding. By further considering the channel condition, a cascaded learning based scalable joint semantic-channel coding algorithm is proposed, which autonomously adapts the semantic coding and channel coding strategies to the specific signal-to-noise ratio (SNR). Simulation results show that SST-V achieves better video reconstruction performance, while significantly reducing the transmission overhead.

Keywords: scalable coding; semantic communication; video transmission

Citation (Format 1): LIU C Y, GUO J J, ZHANG Y M, et al. SST-V: a scalable semantic transmission framework for video [J]. ZTE Communications, 2023, 21(2): 70 - 79. DOI: 10.12142/ZTECOM.202302010

Citation (Format 2): C. Y. Liu, J. J. Guo, Y. M. Zhang, et al., "SST-V: a scalable semantic transmission framework for video," *ZTE Communications*, vol. 21, no. 2, pp. 70 - 79, Mar. 2023. doi: 10.12142/ZTECOM.202302010.

1 Introduction

The wireless communication paradigm is envisioned to shift from connecting things to connecting intelligence, imposing new challenges on the developing sixth generation (6G) communication systems. On the one hand, new intelligent applications, such as the digital twin and the smart city, emerge with a surging number of terminals and explosively increasing data[1], bringing a great burden to existing communication systems. At the same time, to achieve real-time intelligent decision and control, communications are expected to be extremely low-delay and reliable. These challenges become more stringent when it comes to video data, which accounts for more than 80% of Internet traffic[2] and is further rising driven by the demand for ultra-high definition (HD) video. For example, a 1 080P HD video with 50 frames per second requires a bandwidth of 60 - 70 Mbit/s in the advanced H.265 format encoding. As a result, existing coding and transmission strategies, aiming at transmitting every bit, face the dual challenges of bandwidth and delay and are not capable enough of meeting the future demands of ultra-low delay and even real-time video transmission. It is urgent to develop a more efficient video compression and transmission paradigm.

In recent years, the semantic communication driven by artificial intelligence (AI), which is regarded as one of the potential technologies of 6G, has shown great potential due to its superior performance in data compression and transmission. WEAVER and SHANNON[3] divided the communication problems into three levels, namely the technical problem, semantic problem, and effectiveness problem, which corresponds to the definition of syntactic, semantic, and pragmatic in the theory of signs[4]. Based on the syntactic level of information, existing communication systems are developed, aiming at achieving complete and correct transmission of every symbol. Differently, semantic communication systems focus on the semantic level of information and aim at delivering the goal-related

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62293485 and the Fundamental Research Funds for the Central Universities under Grant No. 2022RC18.

parts of messages, which reduces the communication overhead to a great extent. Recently, semantic communications have been widely studied for the text[5], speech[6], and image[7], and demonstrated to be a powerful solution to high-efficiency compression and transmission.

For the video, existing source coding methods, like H.264, take the numerical difference between pixels as the evaluation index of the reconstruction quality, and all pixels are supposed to be transmitted completely. When researchers are coping with the phenomenon of lag due to the limited bandwidth, the clarity of the reconstructed video often gives place to fluency. However, for the human visual perception system, eyes tend to be attracted by some specific parts, e.g., the dynamic information in the foreground of the picture, and are not very sensitive to others like absolute errors between pixels of background. Inspired by this, the trade-off between visual experience and communication overhead can be better achieved in video compression and transmission by preserving the informative and significant parts, i.e., semantic information. Some researchers have investigated the effectiveness of semantic-enabled video compression and transmission. LU et al. proposed the first semantic compression framework for video, named Deep Video Compression (DVC)[8], which implements the video compression modules with convolutional neural networks (CNN) and optimizes them in an end-to-end manner. In Ref. [9], the effect of channel transmission is taken into consideration by jointly designing the semantic and channel coding for video transmission, and the coding rate of semantic information is determined by the entropy model, which is further expanded to frame-level control in Refs. [10] and [11]. Although these works provide efficient solutions to channel-aware semantic transmission, they ignore the change of transmission environments and are suboptimal under dynamic channel conditions.

To improve the efficiency in semantic communications of videos under dynamic channels, this paper proposes a scalable semantic transmission framework for video (SST-V), which achieves adaptive control of the coding rate towards dynamic channels. Specifically, a semantic importance estimation (SIE) module is proposed to evaluate the importance of different semantic features, where the semantic features of higher significance are given higher weights in the successive coding. To improve the efficiency and robustness of semantic transmission, we design a scalable multi-level joint semantic-channel (S-JSC) coding algorithm, where the coding rate of semantic feature is adaptively adjusted according to the corresponding importance level and the specific channel condition. In addition, a cascade-learning-based training strategy is applied for S-JSC, which greatly reduces the training and storage overhead.

The rest of the paper is as follows. Section 2 summarizes the existing video compression standard and semantic transmission methods for video. In Section 3, a basic framework for SST-V is proposed, including an SIE module and an S-JSC coding algorithm. Section 4 gives specific implementation details, simulation results, and performance analysis. Finally, Section 5 concludes the paper.

2 Overview of Video Transmission

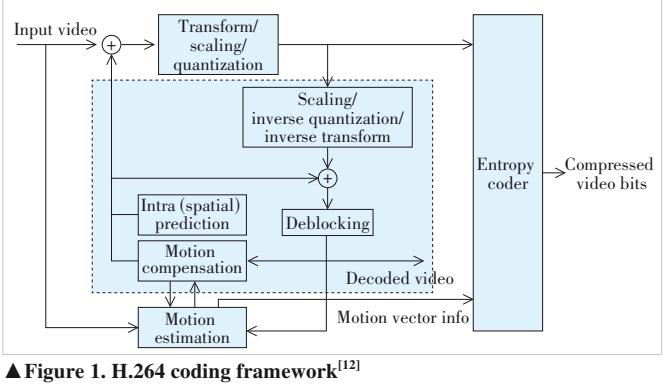
In this section, we introduce the research status and progress of relevant fields, including existing video compression, semantic communication, and semantic transmission of video. Meanwhile, some private opinions are given about problems of current research and possible directions for improvement.

2.1 Existing Video Compression Coding Method

Relevant standards in the field of video coding are mainly formulated by two major organizations: the International Organization for Standardization/the International Electrotechnical Commission (IOS/IEC) and the International Telecommunication Union (ITU-T). The Moving Picture Experts Group (MPEG) of ISO/IEC has formulated the MPEG series of video coding standards for motion image compression. ITU-T has formulated the H.26x series of video coding standards, which is mainly used for low-bit-rate video telephony. There are also some standards that are formulated jointly by IOS/IEC and ITU-T, such as H.264/MPEG-4 part10 and H.265/HEVC.

Jaswant. R. JAIN and Anil. K. JAIN proposed a hybrid coding framework based on block motion compensation (MC) and transform coding such as discrete cosine transform (DCT) in Picture Coding Symposium (PCS) in 1979, which has become the main framework of almost all later video coding standards. The framework is also known as the MC/DCT hybrid coding. The core modules in MC/DCT hybrid coding include predictive coding and transform coding. In order to reduce the complexity of coding and make the operation of video coding easy to execute, each frame is divided into fixed-size blocks first, and then the blocks are compressed and encoded. In addition, there are quantization, entropy coding, and other modules. For example, the H.264 video coding framework is shown in Fig. 1.

In order to reduce the time redundancy among video frames, inter-frame prediction and motion compensation are usually used. MPEG series compression coding standards di-



▲Figure 1. H.264 coding framework[12]

vide different video frames into intra-frames (I-frame), predictive frames (P-frame), and bi-directional interpolated prediction frames (B-frame). For different types of frames, different compression ratios are used to achieve a balance between compression efficiency and video quality. The I-frame is the key frame. In the H.264 standard, video frames of fixed length are divided into a Group of Pictures (GOP) to prevent error propagation among reconstructed frames. The first frame in each GOP, as the key frame, is independently compressed by the Joint Photographic Experts Group (JPEG) or other image encoding methods to maximize the preservation of frame information. The previous I-frame or P-frame is used as a reference frame for the subsequent P-frame. Huffman encoding is performed on the motion vector, and a higher degree of compression is performed on the residual by using approximate JPEG encoding. The B-frame uses the front and back frames for bidirectional interpolation prediction, which has the highest degree of compression, but the decoding complexity and distortion are higher.

However, existing video compression methods have trouble dealing with increasing video data and the reasons are as follows. Firstly, existing standards take minimizing pixel error as the reconstruction goal and ignore the semantic information contained in the video. Secondly, they adopt fixed modular designs in which each module is independent of the others, such as DCT transform and entropy coding. As a result, they cannot obtain an overall performance gain. Benefiting from the development of deep learning, which has a strong nonlinear characterization ability, the latest evolution schemes of mainstream coding methods such as H.265, AVS2, and AVS3 have taken it into consideration to improve the coding performance.

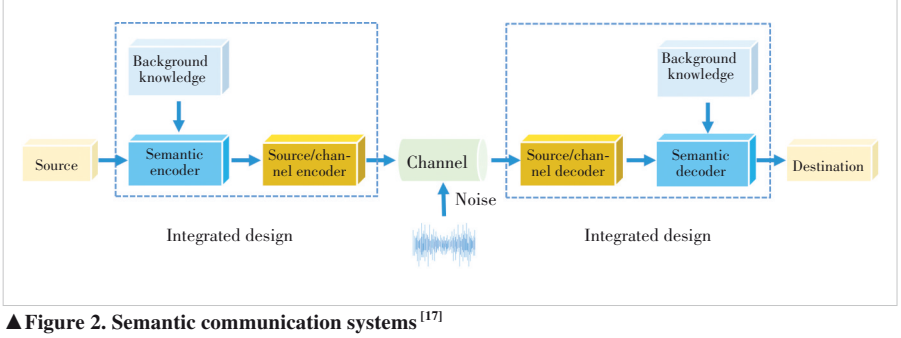
2.2 Video Semantic Communication

The development of semantic information theory[3, 13 – 18] has supported the rapid growth of semantic communications in recent years. Studies on different modalities[19 – 21] have shown that joint source-channel coding can improve the overall performance of the system, and efficiently handle wireless channel fading and interference. Inspired by the joint design, joint semantic-channel coding schemes are widely adopted in semantic communication systems, which combines the semantic representation of the source with the link state of the physical layer.

2.2.1 Semantic Compression for Video

Video semantic compression and reconstruction can be divided into two categories: optimizing the existing video compression framework and extracting key information from videos to be compressed. The optimization of the existing compression framework mainly considers the method with a delay constraint, which means the reference frame is only from previous frames. This makes it more suitable for actual application scenarios like streaming media. Specific modules in the existing video coding framework have been considered to be replaced by neural networks (NN). In Ref. [22], the existing video compression algorithm based on DCT is combined with video frame interpolation based on deep learning. According to the threshold of peak signal-to-noise ratio (PSNR), the encoded data can be selected to provide adjustable compression for residuals. In Ref. [23], four kinds of deformation of attention mechanisms are proposed, which respectively use the I-frame, motion vector, residual error, audio signal, and other mode information for video action recognition. Different types of information are processed in different ways and compared with each other. In 2019, LU et al. proposed an end-to-end deep video compression (DVC) model for the first time[8], which combines the optimization of video compression modules and uses CNN to optimize the network. The CNN network implements an encoder, a decoder, and a motion compensation network and uses a highly nonlinear transformation to represent residuals, which improves compression efficiency. Based on DVC, network modules such as feature prediction, loop filter, and discriminator[24 – 28] are added to further improve compression performance, making end-to-end video compression an important research trend. Multiple frames prediction for learned video compression (MLVC)[24] calculates relative motion using multiple previous frames, thus reducing coding residuals. A deep contextual video compression (DCVC) is proposed in Ref. [25], which uses the feature domain context as a condition and performs conditional coding instead of sub-optimal residuals. Considering the similarity of spatial dependencies, advanced learned video compression (ALVC)[28] predicts the current frame from previous frames without consuming any bits, further reducing coding overhead.

For the second category, key semantic features are extracted from original videos. There are relatively mature attempts for certain types of data sets. Based on the semantic segmentation technology, frames are divided into different semantic units, each of which has a specific spatial arrangement and visual characteristics, and is coded separately[29]. ZHANG et al. extract the semantic features of football video games from the elements of foreground, background, and the relationship between different objects[30]. The encoded sequences of



▲Figure 2. Semantic communication systems[17]

these features are decoded at the receiving end correspondingly, and then fused by a U-net network to generate a complete video. A semantic video conferencing (SVC)[31] network is proposed to extract key points of speakers to realize the semantic transmission of the video conference. CHEN et al. propose a framework for Interactive Face Video Coding (IFVC)[32] where each talking frame is expressed by highly-independent facial features such as mouth motion and eye blinking, achieving superior performance for face videos.

2.2.2 Semantic Transmission for Video

The above works consider video semantic compression schemes under the condition of sufficient bandwidth and ideal channels. Considering practical communication scenarios, new video semantic transmission schemes have been developed to facilitate joint optimization of semantic coding and channel coding. To achieve the balance between video quality and transmission delay in real-time video transmission scenarios, CUI et al.[33] use reinforcement learning (RL) to generate inferencing models based on playback and cache information when network throughput fluctuates. ELGAMAL et al.[34] manage to carry out targeted video coding according to specific downstream tasks in edge computing and cloud computing scenarios. They focus on capturing the target object in the picture. When the target object has a violent change (e.g., vehicle entering or leaving the picture), a new I-frame is selected, and only the I-frame is retrieved under specific tasks to reduce computing and transmission overhead. For surveillance videos[35], only the salient zones are encoded with high resolution, therefore the calculation ability of fog nodes can be reasonably allocated to obtain low delay while maintaining the video quality.

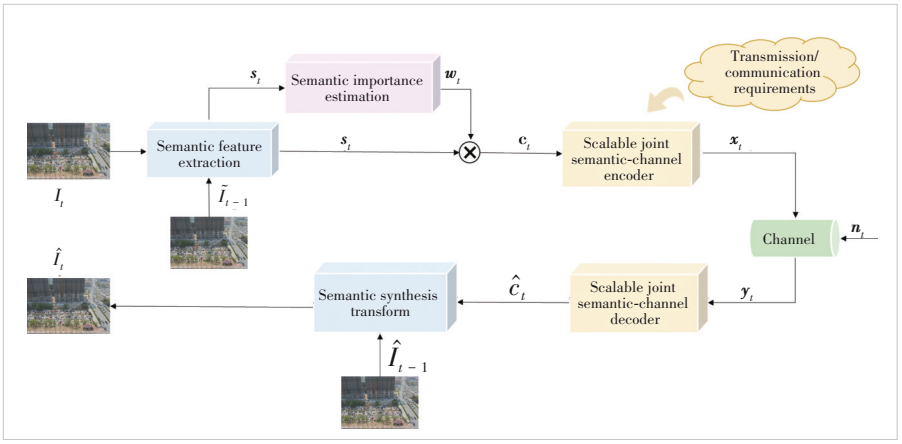
The joint design under specific scenarios considers schemes with a fixed rate, and there is still room for performance improvement for dynamic coding. Therefore, some works have studied variable-length semantic coding methods for the video to further improve the efficiency of semantic encoding. In Ref. [36], the resource allocated to different video frames is determined by the position in GOPs. According to the distance between the video frame and the key frame, a hierarchical learned video compression (HLVC) method is established with three hierarchical quality layers and a recurrent enhancement network. However, this allocation strategy treats the frame as the rate control unit and does not go deep into the level of semantic features, which ignores the importance of different semantic information and is difficult to further compress the content redundancy within frames. In Ref. [9], an entropy model is used to obtain the rate-adaptive transmission strategy, which has several parallel autoencoders with a different number of output channel symbols. Although the proposed framework achieves adaptive control towards the importance of semantic features, which, however, is evaluated with a syntactic information entropy, leading to suboptimal rate control strategies for semantic coding. In Ref. [29], a semantic bit allocation model based on RL is proposed, which aims at improving the rate-semantic-sensing performance by encoding a certain semantic concept. Different features are put into the semantic decoder at different quantization levels according to the allocated resources and then reconstruct the image. These two works achieve a variable length encoding of semantic features, while the bit rates are changed through multiple parallel encoders, which greatly increases the complexity of neural network architecture. In Ref. [37], a rate allocation network is introduced to analyze the semantic information and anti-noise capability of the frame features. Features are coded and transmitted in a descending order of semantic importance according to the mask generated following the rate allocation network, and features with lower importance may be discarded to achieve video transmission of different bit rates. The above works mainly study variable-length coding schemes with different semantic features from the perspective of reconstruction. However, it is still an unsolved problem how to implement a flexible and scalable video semantic transmission scheme when the channel dynamically changes.

3 Proposed Framework of Scalable Semantic Transmission

In this section, we first present the basic framework of scalable video semantic transmission, and then introduce the proposed SIE module and S-JSC coding algorithm. The cascaded training strategy for the proposed system is finally presented.

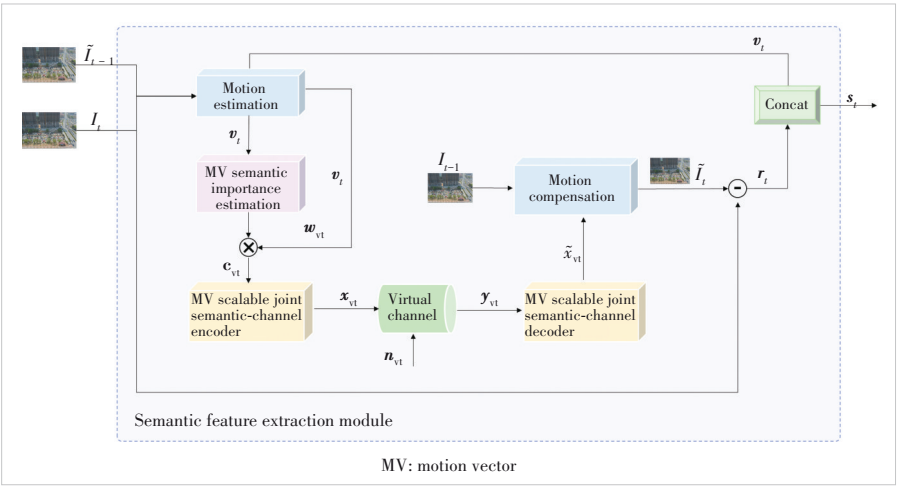
3.1 Proposed Framework

The total framework of the proposed SST-V is shown in Fig. 3. The transmitter consists of a semantic feature extrac-



▲Figure 3. Framework of scalable semantic transmission for video

tion module $S(\cdot)$, a SIE module $A(\cdot)$ and a S-JSC encoder $E(\cdot)$. The receiver mainly consists of a S-JSC decoder $D(\cdot)$ and a semantic synthesis transform $T(\cdot)$. The video $\boldsymbol{I}=\{I_1,I_2,\cdots,I_T\}$ is input into the semantic feature extraction module by frames, while the input image can be expressed as $I_t\in\mathbb{R}^{W\times H}$, where W and H are the width and height of the images, respectively. The semantic feature extraction module compares the semantic information changes between the current image I_t and the previous reference frame $\tilde{I}_{t-1}$ and calculates the semantic information required to reconstruct the video $\boldsymbol{s}_t=S(I_t,I_{t-1})\in\mathbb{R}^{N\times M_W\times M_H}$, where N is the number of the semantic features and $M_W\times M_H$ is the size of each feature map. The weight of different semantic information $\boldsymbol{w}_t=A(\mathbf{s}_t)\in\mathbb{R}^{N\times M_W\times M_H}$ is given by SIE. The semantic information to be transmitted can be expressed as $\boldsymbol{c}_t=\boldsymbol{s}_t\times\boldsymbol{w}_t$. The S-JSC encoder encodes $\boldsymbol{c}_t$ as $\boldsymbol{x}_t=E(\boldsymbol{c}_t)\in C^l$, where l is the number of the transmission symbols, the value of which is chosen among the set $\boldsymbol{L}=\{l_1,l_2,\cdots,l_C\}$ according to the channel state or different communication demands.



▲Figure 4. Framework of semantic feature extraction module

In this paper, we consider both the additive white Gaussian noise (AWGN) channel and the Rayleigh fading channel. As for the AWGN channel, a received sequence can be expressed as $\boldsymbol{y}_t=\boldsymbol{x}_t+\boldsymbol{n}_t\in C^l$. The noise vector $\boldsymbol{n}_t$ consists of independent and equally distributed cyclic symmetric complex Gaussian random variables n_i which follows $n_i\sim\mathcal{CN}(0,\sigma^2),i=1,\cdots,l$, and σ^2 is the average noise power. For Rayleigh fading channels, the single point fading model is considered in this paper and all transmission symbols experience the same channel response. Clarke Model[38] shows that a flat fading channel is composed of several multipath signals under a rich-scattering electromagnetic environment. According to the central limit theorem, both the I-path and Q-path of the channel response can be approximated as Gaussian random processes when the number of paths is large enough. Similar to Refs. [5] and [7] , the received sequence can be expressed as $\boldsymbol{y}_t=h_t\boldsymbol{x}_t+\boldsymbol{n}_t\in C^l$, where $h_t\sim\mathcal{CN}(0,{\sigma_t}^2)$ is a random variable satisfying a cyclic symmetric complex Gaussian distribution.

At the receiver side, the received sequence is decoded as $\hat{\boldsymbol{c}}_t=D(\boldsymbol{y}_t)\in\mathbb{R}^{N\times M_W\times M_H}$ by the S-JSC decoder, which selects different decoder structures according to the number of symbols received. Finally, the decoded sequence of semantic information is transmitted to the semantic synthesis transform module to reconstruct the original video frame $\hat{I}_t\in\mathbb{R}^{W\times H}=T(\boldsymbol{c}_t)$.

3.2 Semantic Feature Extraction

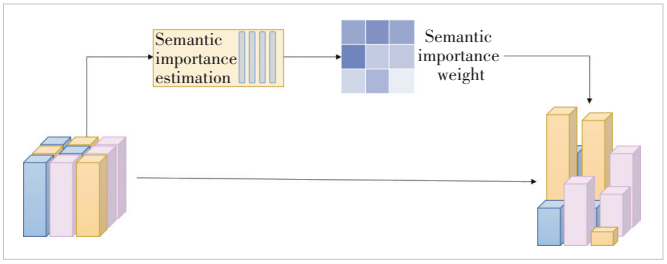
The framework of the semantic feature extraction module $S(\cdot)$ is shown in Fig. 4, which is referred to the end-to-end video compression structure in Ref. [8]. The extracted semantic feature $\boldsymbol{s}_t$ is mainly obtained from the motion vector (MV) $\boldsymbol{v}_t$ and the residual vector (RES) $\boldsymbol{r}_t$.

The current frame I_t and the reference frame $\tilde{I}_{t-1}$ are first fed into the motion estimation module to get the motion vector $\boldsymbol{v}_t$[39]. A semantic feature map of the motion vector with new distribution is obtained in MV SIE and further encoded in an MV S-JSC encoder. In particular, the semantic feature map needs to pass through a virtual channel at the transmitter and then the decoding process is simulated. The decoded semantic sequence $\tilde{\boldsymbol{x}}_{vt}$ is considered to be approximately consistent with the semantic sequence obtained by the receiver. Therefore, the semantic synthesis transform module can predict the received frame from $\tilde{\boldsymbol{x}}_{vt}$ and I_{t-1}. By comparing the error between the simulated reconstructed frame $\tilde{I}_t$ and the original frame I_t, the residual error vector $\boldsymbol{r}_t$ of the current motion vector can be calculated. The encoded motion vector and residual vector are spliced together to get the final sequence $\boldsymbol{s}_t=\{\boldsymbol{v}_t,\boldsymbol{r}_t\}$. Decoded semantic sequence $\hat{\boldsymbol{s}}_t$ after joint semantic channel decoder can be divided into decoded motion vector $\hat{\boldsymbol{v}}_t$ and residual error $\hat{\boldsymbol{r}}_t$ correspondingly. The semantic synthesis firstly carries out motion estimation based on the previous reference frame $\hat{I}_{t-1}$ and $\hat{\boldsymbol{v}}_t$ to generate frame I_{MC-t}, followed by correction using $\hat{\boldsymbol{r}}_t$ to obtain the final reconstructed frame $\hat{I}_t$.

3.3 Semantic Importance Estimation

For video frames, semantic feature vectors indicate what is present in each frame and what has changed compared with the previous frame. However, it is worth noting that these semantic feature vectors have different importance. In the case of street surveillance video, we care more about the movement of cars and pedestrians that dominate the video than the swaying leaves in the wind.

Therefore, we design the SIE module $A(\cdot)$ based on squeeze-and-excitation networks (SE-Net) shown in Fig. 5. Following the semantic features extraction, SIE is firstly used to compre-



▲Figure 5. Semantic importance estimation (SIE) module structure

hensively analyze the relationship between different feature maps to estimate the importance degree of different features, and provide different weights for each feature. Since the semantic information s_t consists of two parts, SIE estimates the semantic importance of v_t and r_t respectively, and the output is $w_t = \{w_{vt}, w_{rt}\}$. Then, the extracted semantic feature is multiplied by the weights to produce a new feature map. On the one hand, more power can be allocated to important information during an actual communication process to reduce the effect of noise. On the other hand, important semantic information needs more strict protection by the S-JSC coding algorithm explained below. When the channel conditions are severe with a low SNR, the correct transmission of important semantic features can be guaranteed with the same number of channel symbols to realize the reconstruction of basic semantic information despite the interference of noise.

3.4 Scalable Joint Semantic-Channel Coding Algorithm

Joint semantic-channel coding is supposed to be used for end-to-end overall optimization, which further enhances the accuracy of the semantic reconstruction of transmitted videos and protects the semantic information obtained in Sections 3.2 and 3.3. In particular, existing video semantic coding methods based on deep learning cannot adjust the code rate flexibly. To solve this problem, the S-JSC coding algorithm is designed, which can adjust the code rate adaptively according to the actual transmission requirements.

According to the training strategy of cascade learning[40], several different source channel coding rates are designed. With the increase in coding level, the output dimension of the S-JSC coding algorithm is continuously reduced while the compression ratio is continuously improved. Higher-level algorithms with fewer symbols manage to maintain the maximum transmission quality within limited resources. Instead of indiscriminately compressing the encoding output of the upper level, semantic features of different importance obtained from SIE are protected with different degrees. The redundancy of semantic information with less importance may be decreased greatly to realize reliable transmission of the most important information, therefore achieving more efficient video semantic transmission.

The coding level is used as the control parameter and input into the S-JSC encoder/decoder together with the information to be encoded. According to the coding level, the scalable autoencoder layer specifies the neural network architecture to change the dimensions of output. The training and storage overhead is greatly reduced with multilevel coding algorithms stored in a serial structure, improving the deployment efficiency of the model.

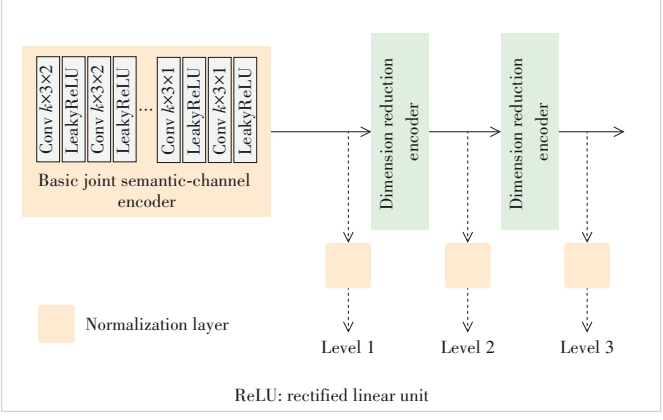
3.5 Training Strategy

During the training process, it is of limited significance for the subsequent layers to participate in the training before the semantic feature extraction module becomes basically stable, due to the relatively complex network architecture and the large correlation between the front and back layers. Therefore, the semantic feature extraction module with relatively stable performance can be obtained by separating the training first. After it is basically stable, SIE and levels of the S-JSC coding algorithm are added successively. Due to the random distribution of parameters of the newly added layer, the semantic feature extraction module is frozen, which makes the subsequent structure converge quickly. When lower levels of S-JSC are trained, there are epochs where all preceding components are updated to achieve an end-to-end gain.

With the gradual addition of the new JSC coding layer, the parameters of the trained S-JSC coding layers and SIE will also be frozen to ensure that the output of the lower coding level is not damaged as much as possible during the training of the higher coding level. The update of parameters will not be carried out in the back propagation, and only the new coding layers will be trained. SIE and the low-level S-JSC encoder layers get the overall gain through the joint training, which improves the anti-interference ability from the perspective of semantic information protection. The high-level S-JSC coding layers mainly further reduce the transmission symbol error rate from the perspective of transmission under insufficient channel carrying capacity.

4 Experiments and Results

In this section, we present the details of training, including data set processing, optimization objectives, and evaluation indicators. Simulation results are analyzed, which proves the ef-



▲Figure 6. Scalable multi-level joint semantic-channel (S-JSC) coding architecture based on cascade learning

fectiveness of the proposed framework.

4.1 Datasets

The proposed system is trained with the Vimeo-90k dataset[41], which is one of the most commonly used video datasets built for evaluating different video processing tasks. The complete dataset is about 82 G in size, and according to Ref. [8], it would take about 7 days to train a single similar semantic feature extraction module using two Titan X GPUs. The dataset contains 89 800 independent short films, each of which consists of 7 frames. During each epoch, according to the time relativity, the relative motion for each frame with respect to its reference frame is considered to be approximate and the coding strategy is therefore similar among the 7 frames of each video clip. Only one frame of each video is selected randomly for our training, together with the previous frame as references. Compared with training each group of the frame and its reference frame, the volume of training details decreases to 1/6 of the original volume, which greatly reduces the time cost while the performance in subsequent validation is basically unchanged.

4.2 Optimization Objective and Metrics

The optimization objective of SST-V is to improve the video reconstruction performance with fewer channel symbols. The entropy of preliminary semantic information s_t from the semantic feature extraction module is supposed to decrease and therefore reduce the difficulty of subsequent module coding. Therefore, the rate-distortion (RD) function is adopted as the loss function, i.e.,

$$\text{loss} = \lambda l_D + l_R = \lambda d(x_t, \hat{x}_t) + (H(v_t) + H(r_t)), \tag{1}$$

where λ is the Lagrange multiplier that represents the tradeoff between bit overhead and video distortion, l_R is the coding bit rate of the semantic feature extraction module, represented by entropies of the moving vector $\boldsymbol{v}_t$ and the residual vector $\boldsymbol{r}_t$, and l_D is the distortion constraint of video reconstruction quality that consists of mean square error (MSE) of the original video frame. Since the reconstruction process includes both motion compensation using motion vector and correction using residual vector, we need to minimize the errors after motion compensation l_{mc} and the overall distortion after reconstruction l_{re}, i.e.,

$$l_d = l_{\text{mc}} + l_{\text{re}} = w*\text{MSE}(I_{\text{MC}-t}, I_t) + \text{MSE}(I_t, \hat{I}_t), \tag{2}$$

where w is weight of the distortion of motion-vector-based reconstruction that decreases with the training process.

PSNR and multi-scale structural similarity index (MS-SSIM) are used to measure the distortion degree of reconstructed frames. PSNR is calculated as:

$$\text{PSNR} = 10\log_{10}(\text{MAX}^2/\text{MSE}), \tag{3}$$

where MSE is the mean square error of the reconstructed image and original image, and MAX represents the maximum pixel value possible for a frame and is set as 1 during the experiment. Compared with PSNR, MS-SSIM is closer to the real perception of human eyes that ranges from 0 to 1, where a higher value indicates lower distortion. It considers that visual distortion is composed of brightness, contrast and structure, and the influence of the distance from the viewer to the image and the density of pixel information on subjective visual experience are further considered. The detailed calculation of MS-SSIM can be found in Refs. [42] and [43].

Similar to k/n in Ref. [21], channel symbols per pixel (CPP) is defined to measure the coding rate of the system. For a fixed resolution $W \times H$, the number of channel input symbols is R, and then CPP is calculated as:

$$\text{CPP} = R/(W \times H). \tag{4}$$

4.3 Simulation Results

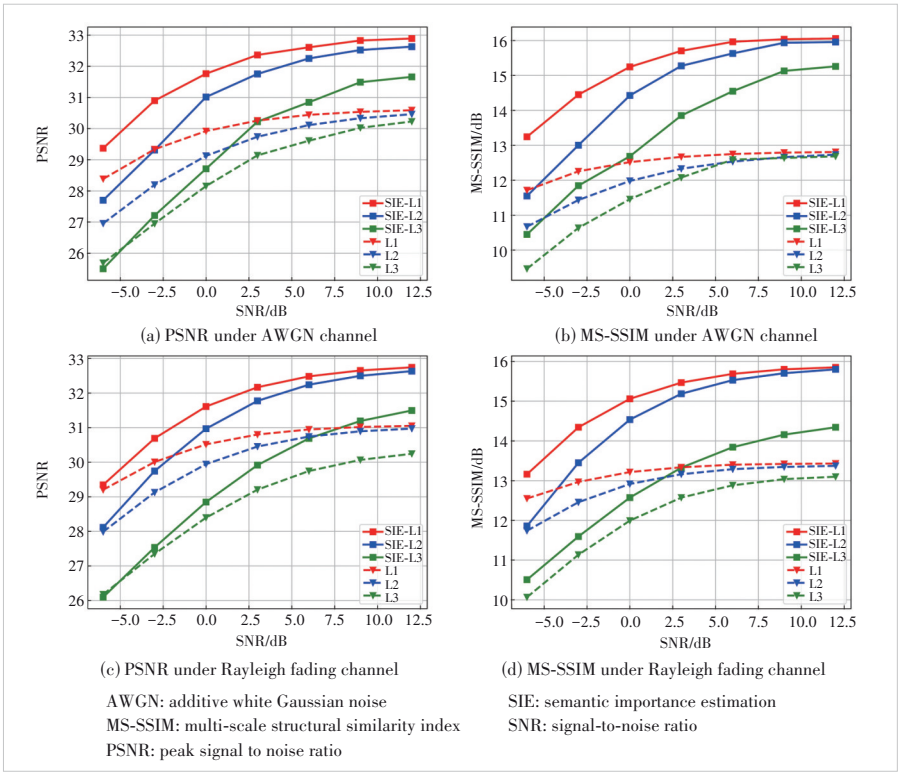
We set up three coding levels, namely Level 1, Level 2, and Level 3, and the value of CPP is 0.110, 0.055, and 0.014, respectively. The ideal channel environment is first considered, where the channel capacity is assumed to always be enough to serve the transmission of all the channel input symbols. Fig. 7 shows the reconstruction performance of the schemes with and without SIE under different coding levels. "SIE-L1" indicates the scheme at Level 1 with SIE, while "L1" means the scheme without SIE.

Since most values of MS-SSIM are distributed densely, we will use both raw values and the form of n dB for better visual effect, which is calculated by

$$\text{MS}-\text{SSIM}(\text{dB}) = -10\log_{10}(1-\text{MS}-\text{SSIM}). \tag{5}$$

Under different channel conditions and metrics, the two curves at the top are both the results of schemes with SIE, which achieve higher performance gain under the metric of MS-SSIM. Note that benefitting from SIE, the scheme at Level 2 performs even better than the non-SIE scheme at Level 1 in most cases. Therefore, SIE helps greatly in extracting important semantic information from the source video, reducing the communication cost significantly. The advantage of SIE is also shown in Fig. 8, which gives several visual results of the schemes with or without SIE at 0 dB with Level 3 under the ideal channel.

However, in the practical communication system, the channel capacity is not always sufficient and deteriorates severely under bad channel conditions. Hence, the proposed schemes are also evaluated under capacity-limited cases. During the test process, we make SNR vary uniformly from 0 dB to 10 dB. The average PSNR and MS-SSIM of different schemes are shown in Table 1, where schemes 1 and 2 perform better than schemes 3 and 4 under non-ideal channels, respectively, which further demonstrates the effectiveness of the proposed



▲Figure 7. Reconstruction performance of the schemes with or without SIE at different coding levels

▲Figure 8. Examples of the reconstructed frames of the schemes with or without semantic importance estimation (SIE)

SIE. Scheme 3 performs worse than scheme 4 even if the former uses more symbols for transmission. The reason is that all the semantic information is transmitted without considering the variation of channel capacity and the important information is deteriorated with higher probability, exposing the limitation of the fixed level coding scheme.

For the proposed scalable multilevel coding scheme, the coding level increases for lower SNR. Level 3, Level 2 and Level 1 are selected respectively for channel conditions ranging in [0,3), [3,6) and [6,10]. The number of three levels is assumed to be the maximum number of symbols that can be accurately transmitted under the corresponding channel conditions. For the proposed SST-V, each frame can automatically switch to a different coding model according to the SNR, while the non-scalable baselines (schemes 1, 2, 3 and 4) use a fixed encoding level at all SNRs.

As shown in Table 1, compared with schemes 3 and 4, the proposed SST-V with both SIE and S-JSC, i.e., scheme 6, achieves performance gains of 2.3 dB and 4.6 dB in terms of PSNR and MS-SSIM, respectively. It shows that when the channel capacity is limited, the SST-V can adapt to the dynamic channel environment, significantly improving the transmission efficiency. Note that schemes 5 and 6 further evaluate the effectiveness of SIE under the scalable multilevel coding rate. With the proposed SIE, scheme 6 achieves gains of 1.3 dB in PSNR and 2.7 dB in MS-SSIM. It proves that SIE helps the SST-V focus on the semantic information of higher importance, and hence improves the reconstruction performance under practical dynamic channels.

▼Table 1. PSNR and MS-SSIM of different schemes

Number	Scheme	PSNR	MS-SSIM/dB
1	Without SIE, fixed Level 1 (L1)	23.937 6	6.704 7
2	Without SIE, fixed Level 2 (L2)	27.307 8	8.743 6
3	With SIE, fixed Level 1 (SIE-L1)	26.099 3	7.454 1
4	With SIE, fixed Level 2 (SIE-L2)	28.900 34	9.754 9
5	Scalable multilevel coding without SIE	29.935 9	11.682 3
6 (SST-V)	Scalable multilevel coding with SIE	31.190 78	14.349 9

MS-SSIM: multi-scale structural similarity index
PSNR: peak signal-to-noise ratio
SIE: semantic importance estimation
SST-V: scalable semantic transmission framework for video

5 Conclusions

In this paper, we discuss the video transmission problem in the future 6G mobile communication scenarios and review the existing video coding and semantic-based video coding transmission methods. To achieve efficient and robust video transmission under dynamic channel conditions, this paper proposes a scalable semantic transmission framework for video, namely SST-V. Besides semantic information extraction, SST-V estimates the importance of different semantic features with the proposed SIE, and obtains a more compact and robust rep-

resentation of semantic information. An S-JSC coding algorithm based on cascading learning is designed, where the coding rate can be adjusted adaptively according to dynamic channel states. The simulation results show that SST-V has better video reconstruction performance in terms of PSNR and MS-SSIM compared with the baseline schemes, and provides a more efficient solution to video transmission under bandwidth constraints.

References

[1] TU Y, CHEN W. A deep learning-based semantic communication system [J]. Mobile communications, 2021, 45(4): 91-94. DOI: 10.3969/j. issn. 1006-1010.2021.04.015
[2] CISCO. 2020 global networking trends report [EB/OL]. (2019-11-17) [2023-04-01]. https://www. cisco. com/c/dam/m/en_us/solutions/enterprise-networks/networking-report/files/GLBL-ENG_NB-06_0_NA_RPT_PDF_MOFU-no-NetworkingTrendsReport-NB_rpten018612_5.pdf
[3] WARREN W, SHANNON C E. Recent contributions to the mathematical theory of communication [EB/OL]. [2023-02-01]. http://www. sietmanagement. fr/wp-content/uploads/2016/04/Weaver1949.pdf
[4] MORRIS C W. Foundations of the theory of signs [M]. Chicago, USA: The University of Chicago Press, 1938
[5] XIE H Q, QIN Z J, LI G Y, et al. Deep learning enabled semantic communication systems [J]. IEEE transactions on signal processing, 2021, 69: 2663 - 2675. DOI: 10.1109/TSP.2021.3071210
[6] WEI H, XU W J, WANG F Y, et al. SemAudio: semantic-aware streaming communications for real-time audio transmission [C]//IEEE Global Communications Conference. IEEE, 2022: 3965 - 3970. DOI: 10.1109/GLOBECOM48099.2022.10001043
[7] XU W J, ZHANG Y M, WANG F Y, et al. Semantic communication for the Internet of vehicles: a multiuser cooperative approach [J]. IEEE vehicular technology magazine, 2023, 18(1): 100 - 109. DOI: 10.1109/MVT.2022.3227723
[8] LU G, OUYANG W L, XU D, et al. DVC: an end-to-end deep video compression framework [C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020: 10998 - 11007. DOI: 10.1109/CVPR.2019.01126
[9] WANG S X, DAI J C, LIANG Z J, et al. Wireless deep video semantic transmission [J]. IEEE journal on selected areas in communications, 2023, 41(1): 214 - 229. DOI: 10.1109/JSAC.2022.3221977
[10] TUNG T Y, GÜNDÜZ D. DeepWiVe: deep-learning-aided wireless video transmission [J]. IEEE journal on selected areas in communications, 2022, 40(9): 2570 - 2583. DOI: 10.1109/JSAC.2022.3191354
[11] HUANG B W, YAN X, ZHOU J J, et al. CSMCNet: scalable video compressive sensing reconstruction with interpretable motion estimation [EB/OL]. (2021-08-03) [2023-02-01]. arXiv: 2108.01522. https://arxiv.org/abs/2108.01522
[12] WIEGAND T, SULLIVAN G J, BJONTEGAARD G, et al. Overview of the H.264/AVC video coding standard [J]. IEEE transactions on circuits and systems for video technology, 2003, 13(7): 560 - 576. DOI: 10.1109/TCSVT.2003.815165
[13] CARNAP R, BAR-HILLEL Y. An outline of a theory of semantic information [EB/OL]. [2023-02-01]. https://courses. cs. tau. ac. il/0368-4341/shared/Papers/CARNAP-HILLEL.pdf
[14] BAR-HILLEL Y, CARNAP R. Semantic information [J]. The British journal for the philosophy of science, 1953, 4(14): 147 - 157. DOI: 10.1093/bjps/iv.14.147
[15] FLORIDI L. Outline of a theory of strongly semantic information [J]. Minds and machines, 2004, 14(2): 197 - 221. DOI: 10.1023/B: MIND.0000021684.50925.c9
[16] KOLCHINSKY A, WOLPERT D H. Semantic information, autonomous agency and non-equilibrium statistical physics [J]. Interface focus, 2018, 8(6): 20180041. DOI: 10.1098/rsfs.2018.0041
[17] ZHANG P, XU W J, GAO H, et al. Toward wisdom-evolutionary and primitive-concise 6G: a new paradigm of semantic communication networks [J]. Engineering, 2022, 8: 60 - 73. DOI: 10.1016/j.eng.2021.11.003
[18] ZHONG Y X. A theory of semantic information [J]. China communications, 2017, 14(1): 1 - 17. DOI: 10.1109/CC.2017.7839754
[19] RAO M, FARSAD N, GOLDSMITH A. Variable length joint source-channel coding of text using deep neural networks [C]//IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC). IEEE, 2018: 1 - 5. DOI: 10.1109/SPAWC.2018.8445924
[20] BOURTSOULATZE E, BURTH KURKA D, GÜNDÜZ D. Deep joint source-channel coding for wireless image transmission [J]. IEEE transactions on cognitive communications and networking, 2019, 5(3): 567 - 579. DOI: 10.1109/TCCN.2019.2919300
[21] KURKA D B, GÜNDÜZ D. DeepJSCC-f: deep joint source-channel coding of images with feedback [J]. IEEE journal on selected areas in information theory, 2020, 1(1): 178 - 193. DOI: 10.1109/JSAIT.2020.2987203
[22] JALALPOUR Y, WANG L Y, FENG W C, et al. FID: frame interpolation and DCT-based video compression [C]//IEEE International Symposium on Multimedia (ISM). IEEE, 2021: 218 - 221. DOI: 10.1109/ISM.2020.00045
[23] CHEN J W, HO C M. MM-ViT: multi-modal video transformer for compressed video action recognition [C]//IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE, 2022: 786 - 797. DOI: 10.1109/WACV51458.2022.00086
[24] LIN J P, LIU D, LI H Q, et al. M-LVC: multiple frames prediction for learned video compression [C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020: 3543 - 3551. DOI: 10.1109/CVPR42600.2020.00360
[25] LI J, LI B, LU Y. Deep contextual video compression [EB/OL]. [2023-04-01]. https://arxiv.org/pdf/2109.15047.pdf
[26] LIU C, SUN H M, ZENG X Y, et al. Learned video compression with residual prediction and feature-aided loop filter [C]//IEEE International Conference on Image Processing (ICIP). IEEE, 2022: 1321 - 1325. DOI: 10.1109/ICIP46576.2022.9897989
[27] ZHANG S P, MRAK M, HERRANZ L, et al. DVC-P: deep video compression with perceptual optimizations [C]//Proceedings of 2021 International Conference on Visual Communications and Image Processing (VCIP). IEEE, 2022: 1 - 5. DOI: 10.1109/VCIP53242.2021.9675350
[28] YANG R, TIMOFTE R, VAN GOOL L. Advancing learned video compression with In-loop frame prediction [J]. IEEE transactions on circuits and systems for video technology, 2023, 33(5): 2410 - 2423. DOI: 10.1109/TCSVT.2022.3222418
[29] HUANG D L, GAO F F, TAO X M, et al. Toward semantic communications: deep learning-based image semantic coding [J]. IEEE journal on selected areas in communications, 2023, 41(1): 55 - 71. DOI: 10.1109/JSAC.2022.3221999
[30] DUAN Y P, LI M Z, WEN L J, et al. From object-attribute-relation semantic representation to video generation: a multiple variational autoencoder approach [C]//Proceedings of 2022 IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP). IEEE, 2022: 1 - 6. DOI: 10.1109/MLSP55214.2022.9943394
[31] JIANG P W, WEN C K, JIN S, et al. Wireless semantic communications for video conferencing [J]. IEEE journal on selected areas in communications, 2023, 41(1): 230 - 244. DOI: 10.1109/JSAC.2022.3221968
[32] CHEN B, WANG Z, LI B, et al. Interactive face video coding: a generative compression framework [EB/OL]. [2023-02-20]. https://arxiv. org/abs/2302.09919
[33] CUI L Z, SU D Y, YANG S, et al. TCLiVi: transmission control in live video streaming based on deep reinforcement learning [J]. IEEE transactions on Multimedia, 2020, 23: 651-663. DOI: 10.1109/TMM.2020.2985631
[34] ELGAMAL T, SHI S, GUPTA V, et al. SiEVE: semantically encoded video analytics on edge and cloud [C]//Proceedings of 2020 IEEE 40th International Conference on Distributed Computing Systems (ICDCS). IEEE, 2021: 1383 - 1388. DOI: 10.1109/ICDCS47774.2020.00182
[35] WANG Y Q, XU J C, JI W. A feature-based video transmission framework for visual IoT in fog computing systems [C]//Proceedings of 2019 ACM/IEEE Symposium on Architectures for Networking and Communications Systems (ANCS). IEEE, 2019: 1 - 8. DOI: 10.1109/ANCS.2019.8901872
[36] YANG R, MENTZER F, VAN GOOL L, et al. Learning for video compression with hierarchical quality and recurrent enhancement [C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

IEEE, 2020: 6627 – 6636. DOI: 10.1109/CVPR42600.2020.00666
[37] ZHANG B, QIN Z, LI Y. Semantic communications with variable-length coding for extended reality [EB/OL]. [2023-03-11]. https://arxiv.org/abs/2302.08645.
[38] RAPPAPORT T S. Wireless communications: principles and practice [M]. Upper Saddle River, USA: Prentice Hall PTR, 1996
[39] RANJAN A, BLACK M J. Optical flow estimation using a spatial pyramid network [C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017: 2720 – 2729. DOI: 10.1109/CVPR.2017.291
[40] MARQUEZ E S, HARE J S, NIRANJAN M. Deep cascade learning [J]. IEEE transactions on neural networks and learning systems, 2018, 29(11): 5475 – 5485. DOI: 10.1109/TNNLS.2018.2805098
[41] XUE T F, CHEN B A, WU J J, et al. Video enhancement with task-oriented flow [J]. International journal of computer vision, 2019, 127(8): 1106 – 1125. DOI: 10.1007/s11263-018-01144-2
[42] 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
[43] WANG Z, SIMONCELLI E P, BOVIK A C. Multiscale structural similarity for image quality assessment [C]//The 37th Asilomar Conference on Signals, Systems & Computers. IEEE, 2004: 1398 – 1402. DOI: 10.1109/ACSSC.2003.1292216

Biographies

LIU Chenyao received her BE degree from the School of Information and Communication Engineering, Beijing University of Posts and Telecommunication (BUPT), China in 2022. She is currently pursuing her PhD degree at the School of Artificial Intelligence, BUPT. Her research interests include semantic communication, video coding, and machine learning.

GUO Jiejie is currently pursuing her BE degree from the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, China. Her research interests include semantic communication, video coding, and artificial intelligence.

ZHANG Yimeng received her BE degree from the School of Information and Communication Engineering, Beijing University of Posts and Telecommunication (BUPT), China in 2018. She is currently pursuing her PhD degree at the School of Artificial Intelligence, BUPT. Her research interests include semantic communication and intelligent resource allocation in emerging wireless applications. She is a graduate student Member of IEEE.

XU Wenjun (wjxu@bupt.edu.cn) is a professor with the State Key Laboratory of Network and Switching Technology, Beijing University of Posts and Telecommunications, China, and with Peng Cheng Laboratory, China. He received his PhD degree from Beijing University of Posts and Telecommunications in 2008. His research interests include artificial intelligence-driven networks, semantic communications, unmanned aerial vehicle communications and networks, and green communications and networking. He is an editor of *China Communications* and a senior member of IEEE.

LIU Yiming received her BE degree in communication engineering from Shanghai University, China in 2014, and PhD degree in information and communication engineering from Beijing University of Posts and Telecommunications (BUPT), China in 2019. She was a visiting PhD student with The University of British Columbia, Canada in 2017 and 2018. She is currently an associate researcher with the School of Information and Communication Engineering, BUPT. Her research interests include next-generation wireless networks, semantic communication, edge intelligence, blockchain, and the distributed ledger technology.