Next Generation Semantic and Spatial Joint Perception —— Neural Metric–Semantic Understanding

ZHU Fang[1,2]

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

DOI: 10.12142/ZTECOM.202101008

http: //kns.cnki.net/kcms/detail/34.1294.TN.20210218.1753.002.html, published online February 19, 2021

Manuscript received: 2020-12-25

Abstract: Efficient perception of the real world is a long-standing effort of computer vision. Modern visual computing techniques have succeeded in attaching semantic labels to thousands of daily objects and reconstructing dense depth maps of complex scenes. However, simultaneous semantic and spatial joint perception, so-called dense 3D semantic mapping, estimating the 3D geometry of a scene and attaching semantic labels to the geometry, remains a challenging problem that, if solved, would make structured vision understanding and editing more widely accessible. Concurrently, progress in computer vision and machine learning has motivated us to pursue the capability of understanding and digitally reconstructing the surrounding world. Neural metric-semantic understanding is a new and rapidly emerging field that combines differentiable machine learning techniques with physical knowledge from computer vision, e.g., the integration of visual-inertial simultaneous localization and mapping (SLAM), mesh reconstruction, and semantic understanding. In this paper, we attempt to summarize the recent trends and applications of neural metric-semantic understanding. Starting with an overview of the underlying computer vision and machine learning concepts, we discuss critical aspects of such perception approaches. Specifically, our emphasis is on fully leveraging the joint semantic and 3D information. Later on, many important applications of the perception capability such as novel view synthesis and semantic augmented reality (AR) contents manipulation are also presented. Finally, we conclude with a discussion of the technical implications of the technology under a 5G edge computing scenario.

Keywords: visual computing; semantic and spatial joint perception; dense 3D semantic mapping; neural metric-semantic understanding

Citation (IEEE Format): F. Zhu, "Next generation semantic and spatial joint perception—neural metric-semantic understanding," *ZTE Communications*, vol. 19, no. 1, pp. 61 – 71, Mar. 2021. doi: 10.12142/ZTECOM. 202101008.

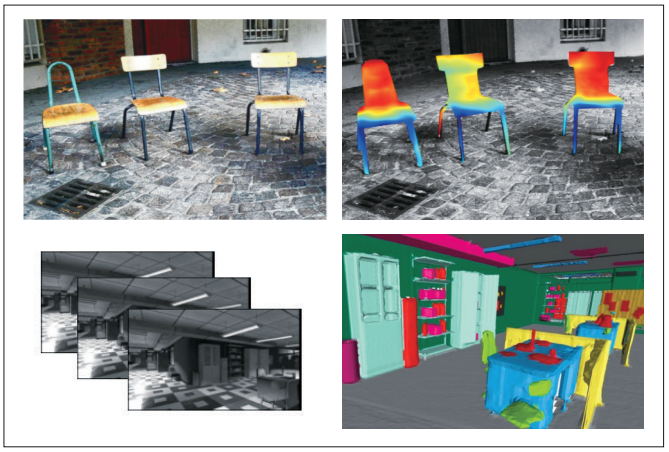
1 Introduction

The perception of the real world in a meaningful reconstructive way has been one of the primary driving forces for the development of sophisticated computer vision techniques. The semantic and spatial joint perception of a variety of scenes is shown in **Fig. 1**. Computer vision approaches span a range from real-time mapping, which enables the latest generation of robots, to sophisticated semantic identification for the meaningfully structured information in various big data applications. In both cases, one of the main bottlenecks is the exact and consistent context understanding in terms of occlusion, view-angle, and illumination conditions, i.e., despite of the noticeable progress in fine-grained semantic scene understanding tasks like detection and instance segmentation, computers still perform unsatisfactorily on visually understanding humans in crowded scenes. Concurrently, powerful consistent context understanding models have emerged in the computer vision and machine learning communities. The seminal works related to semantic and spatial joint perception, the so-called dense 3D semantic mapping framework by HERMANS et al.[1], have evolved in recent years into joint volumetric 3D reconstruction and semantic segmentation formulas for both the unmanned system and the human-involved virtual/augmented reality (VR/AR) immersive experience. Here, the synthesis of more plausible depth in parts of the scene or more reliable semantic image classification can be achieved by jointly optimizing geometry and semantics in 3D. Very recently, such an area has been explored as "metric-semantic understanding". One of the first publications that used

the term metric-semantic understanding is Kimera[2]. It enables machines to learn to perceive their surroundings by combining the-state-of-the-art geometric and semantic understanding into a modern perception way. Furthermore, the authors also argue that the semantic information based on the geometric information provides the ideal level of abstraction to provide humans with models of the environment that are easy to understand. Instead of implicitly combining the geometry and semantic segmentation of 3D, a variety of other methods more explicitly follow this notion of collaboration to exploit components of the perception pipeline.

While classical computer vision starts from the affine imaging of the physical world to addressing the geometrical consistency by modeling, for example, the camera's viewpoint, odometry, and depth map properties, machine learning comes from an end-to-end trainable (differentiable) and statistical perspective. It is a well-known fact that the differentiable machine learning technique can capture more complex dependencies and achieve a high level of expressiveness, while, if used only, cannot be metric or explicitly follow the strict consistency behind the physical world. To this end, the quality of mainly traditional computer vision-based dense 3D semantic mapping relies on the physical correctness of the employed models. Direct joint estimation of geometry and semantics in a multi-view 3D reconstruction setting, which implicitly combines the geometry and semantic information in the scenes, is hard and error-prone and leads to artifacts in the reconstructed map. Thus, the classic computer vision-based geometry reconstructions suffer from not only classical issues, such as poorly textured areas, repetitive patterns, and occlusions, but also several additional challenges, such as higher noise level, and, often, the presence of shake and motion blur. To this end, traditional metric-semantic understanding methods try to overcome these issues by using heuristic regularization, like convex anisotropic regularizers, to combine captured imagery. But in the complex scenery, these methods require thousands of iterations for convergence or are unable to fully capture the complex semantic and geometric dependencies behind them. Neural metric-semantic understanding brings the promise of addressing both geometry reconstruction and fusion of geometry and semantic information by using deep networks to learn complex mappings from captured images to 3D semantic maps. The underlying principle is to combine the differentiable machine learning techniques with physical knowledge from computer vision to yield new and powerful algorithms for semantic and spatial collaborative perception.

Neural metric-semantic understanding does not yet have a clear definition in the literature. Here, we define neural metric-semantic understanding as: deep image or video semantic & spatial collaborative perception approaches and also submodules that enable the explicit or implicit fusion of semantic and geometric context properties of the scene, such as deep convolutional neural networks in volumetric space for 3D semantic segmentation, incorporation of conventional multi-view stereo concepts within a deep learning framework, fine-tuning of the deep network by using the extracted geometric constraints, and a representation of semantics as an invariant scene for medium-term continuous tracking of large scale 3D scanning.



▲**Figure 1. Semantic and spatial joint perception of a variety of scenes**[2-3]

This paper defines the components of the semantic and spatial collaborative perception pipeline and exploits the different directions of neural metric-semantic understanding formulations, embedded in corresponding components. One central scheme around which we structure this paper is the combination of computer vision imaging principles and learning-based primitives to yield new and powerful algorithms for visual content's consistent understanding, since consistency in the real-world understanding is essential for many media editing and structural data indexing applications. We start by discussing previous explorations' fundamental concepts and components of metric-semantic understanding, which are prerequisites for the semantic and spatial collaborative perception pipeline. Afterwards, we discuss critical aspects of emerging neural-based metric-semantic understanding approaches, fusions of learning-based primitives and affine imaging principles, such as type of fusion, how the fusion is provided, which components of the metric-semantic understanding pipeline are learned, and explicit v.s. implicit fusion. Following, we discuss the panorama of applications that is enabled by semantic and spatial collaborative perception. The applications range from relighting, novel view synthesis, to the manipulation of semantic contents for augmented reality (AR). The semantic manipulation of AR contents, achieving natural interaction between the virtual and real world and finally facilitating natural interaction between "digital twins" and the real world, has many technical implications on the evolving storage-computing network, especially when instant response computing and privacy preserving strategies can be carried out with the help of edge computing based on 5G. We then conclude with these implications.

2 Related Surveys

Metric-semantic understanding, sometimes called "dense 3D semantic mapping", has been continuously studied in the literature, such as Ref. [2] and Refs. [4 – 8]. It includes robot perception and mixed reality. The perceptional understanding using classic computer vision or with some convolutional neural networks (CNNs) as classification assistance has been studied extensively. The thorough analysis survey[9] of such classical computer vision methods, for the implicit combination of the geometry and semantic segmentation of 3D, focuses on specific heuristic regularization, such as surface normal directions[10] and special treatment for highly reflective objects[11]. Recent explorations regarding explicitly semantic and spatial collaborative perception through the components of the perception pipeline, with the emerging machine learning techniques, have also been discussed in Refs. [12 – 15]. Recent reports, like Refs. [16 – 19], discuss various applications with the help of metric-semantic understanding techniques, such as novel view synthesis, relighting, and semantic AR contents manipulation. However, none of the above reports or literature provides a structured or comprehensive look into the new and rapidly emerging field, neural metric-semantic understanding, which combines differentiable machine learning techniques with physical knowledge. Such a comprehensive approach, especially linking clues from classic computer vision to the "new" neural assistance, is critical, since the "next generation" semantic and spatial collaborative perception can reach new heights in the performance of these tasks, which motivates us to pursue the modern computer vision capability of understanding and digitally reconstructing the surrounding world.

3 Theoretical Fundamentals

In this section, we discuss the theoretical fundamentals of working in the semantic and spatial collaborative perception space. First, we discuss dense depth map formation models in computer vision, followed by the classic methods of high-quality 3D scanning of large-scale scenes. Next, we discuss approaches to semantic generative models in deep learning. In the end, we discuss the core principles of volumetric semantic 3D reconstruction.

3.1 Dense Depth Map Formation

Classical computer vision methods approximate the reverse prediction process of image formation in the real world. Light sources emit photons that interact with the objects in the scene, as a function of their geometry and material properties, before being recorded by multiple cameras with overlapping views. This process is known as dense depth estimation. Early passive stereo methods, referred to as an in-depth analysis in Ref. [20], relied on at least two recorded frames based on the known camera geometry to extract stereo correspondence, the so-called dense disparity map. Among them, some multi-view stereo methods use multi-valued, voxel-based, or layer-based presentations, while most stereo correspondence methods compute a univalued disparity function $d(x, y)$ with respect to a reference image. The central element to methods that produce a univalued disparity map $d(x, y)$ is the concept of a disparity space (x, y, d). The term disparity describes the difference in the location of corresponding features seen by the left and right eyes. The correspondence between a pixel (x, y) in reference image r and a pixel (x', y') in matching image m is then given by Eq. (1). And the common steps in the stereo algorithms generally include matching cost computation, support aggregation, disparity computation, and disparity optimization. The actual sequence of steps taken depends on the specific algorithm.

$$x' = x + sd(x,y), \qquad y' = y. \tag{1}$$

Passive stereo matching algorithms work well on textured scenes but require demanding computation. Later on, active stereo methods (e.g, Kinect), which triangulate correspondences between a structured active illumination and a camera, have raised a lot of interest. While unstructured surfaces are no longer a problem, the lateral resolution of the active stereo-only methods is limited by the resolution of the projection system under the constraint of size or power. Currently, accurate real-time dense depth estimation is mostly fulfilled with the fusion of sensors, which ultimately improves speed, robustness and quality. A thorough re-inspection regarding the classical paradigm and the fusion between the time of flight (ToF) and stereo, can refer to Ref. [21]. To exploit the complementary strengths, accurate but sparse active range measurements and the ambiguous but dense passive stereo information must be fused under the principle described in Eq. (2) below.

$$E(d) = w_{\text{stereo}}E_{\text{Stereo}}(d) + w_{\text{ToF}}E_{\text{ToF}}(d|d_{\text{ToF}}) + R_{\text{smooth}} + R_{\text{temp}}, \tag{2}$$

where w_{stereo} and w_{ToF} represent confidence/weights, E represents the objective energy to be minimized, and R represents the regularizer.

Different optimization strategies can refer to variably concrete formulas corresponding to the principle described in Eq. (2), such as the local method in Eq. (3) and the variational framework in Eq. (4).

$$E(z_i) = wE_{\text{ToF}}(z_i|z_i^{\text{ToF}}) + (1-w)E_{\text{stereo}}(z), \tag{3}$$

$$E_{\text{data}}(u) := \int_{\Omega} \chi_{\text{ToF}}(x)\rho_{\text{ToF}}(u(x)) + \chi_{\text{Stereo}}(x)\rho_{\text{Stereo}}(u(x))dx. \tag{4}$$

In Eq. (4), ρ represents the local term for penalizing the deviation from the ToF or stereo data, and X represents spatial

indicator functions for valid/trusted ToF/stereo.

3.2 3D Scanning of Large-Scale Scenes

Given the accurate dense depth map of the observed view, high-quality consistent 3D scanning of large-scale scenes is the next key step to the geometric and photometric registration between the virtual and real world. The most important tasks under the objective are estimating globally optimized poses, robust tracking with recovery from gross tracking failures, and re-estimating the 3D model to ensure global consistency, as mentioned by DAI et al.[22]. The core of the above tasks is a robust pose estimation strategy, which globally optimizes the camera trajectory per frame, considering the complete history of the single view depth and image input in an efficient local-to-global hierarchical optimization framework, as described in Refs. [22 – 24]. While each has trade-offs, global optimization methods based on implicit bundle adjustment (BA) are the de facto methods for the highest quality reconstructions. Finally, the optimization for both dense photometric and geometric alignment is based on the energy illustrated in Eq. (5):

$$E_{icp} = \sum_{k} ((v^k - \exp(\hat{\xi}) T v_t^k) \cdot n^k)^2,$$

$$E_{rgb} = \sum_{u \in \Omega} (I(u, C_t^l) - I(\pi(K \exp(\hat{\xi}) T p(u, D_t^l)), \hat{C}_{t-1}))^2,$$

$$E_{track} = E_{icp} + w_{rgb} E_{rgb}, \tag{5}$$

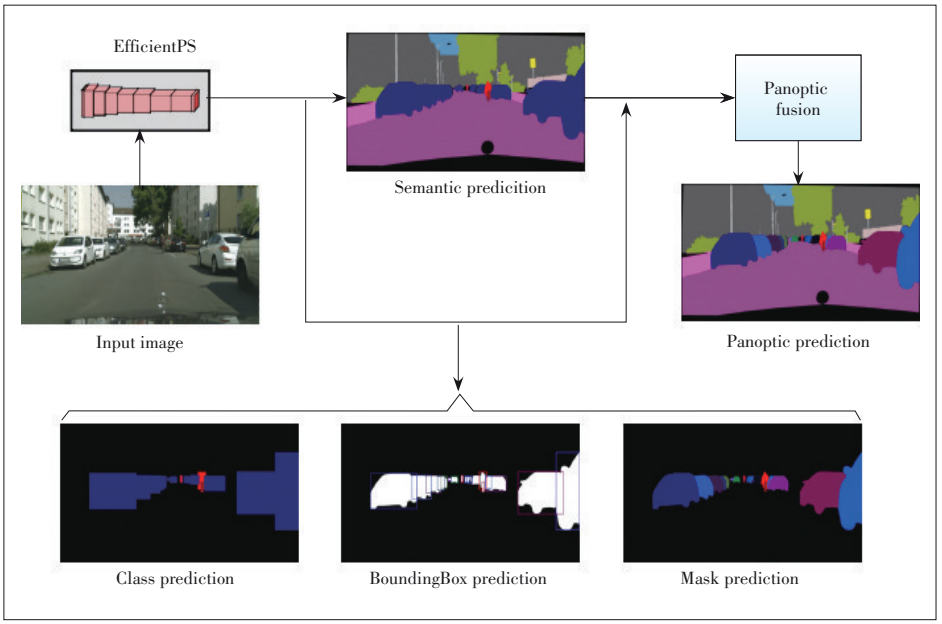
where v^k represents the back-projection of the k-th vertex and n^k is the corresponding normal; D represents the live depth map and C represents the live color image; ξ is the motion parameter and $\exp(\xi)$ is the matrix exponential that maps a member of the Lie Algebra se3 to a member of the corresponding Lie group SE3; T is the current estimate of the transformation from the previous camera pose to the current one; E represents the cost function that needs to be minimized and w represents manually defined weights.

3.3 Semantic Understanding

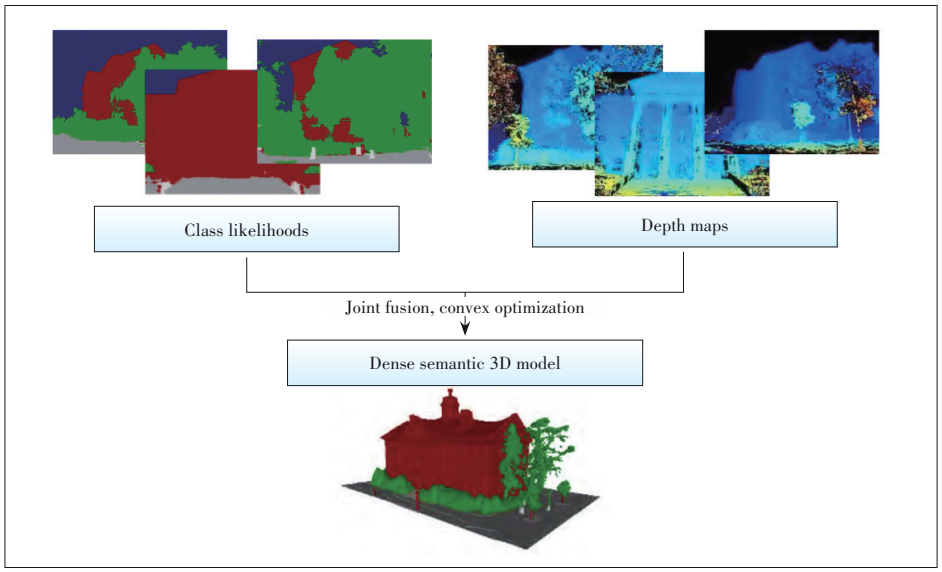
Besides the geometric and photometric registration following the above methods, semantic generative models assist in semantic content registration of the corresponding large-scale scenes. Such scene comprehension, which necessitates recognizing instances of scene participants along with general scene semantics, can be addressed by the panoptic segmentation task with corresponding semantic generative models such as those in Refs. [25 – 26]. Such semantic generative models generally need a deep neural network (e.g, Feature Pyramid Network) as a backbone to efficiently encode and fuse semantically rich multi-scale features, which is followed by a panoptic head network to extract coherently understandable visual scenes at both the fundamental pixel level and distinctive object instance level, as shown in **Fig. 2**. The model predicts four outputs: semantics prediction from the semantic head, class, bounding box, and mask prediction from the instance head. All the aforementioned predictions are then fused in the panoptic fusion module to yield the final panoptic segmentation output. Moreover, advances in the state-of-the-art deep learning methods continually boost the performance of these tasks to new heights.

3.4 Volumetric Semantic 3D Reconstruction

With the above programs, depth maps and pixel-wise semantic classification scores are achieved as inputs to the final objective, the semantic understanding of 3D environments. The core processing will be carried by the volumetric semantic reconstruction, which is cast as a volumetric fusion of depth maps and pixel-wise semantic classification scores. In practical applications, 3D reconstruction systems or semantic segmentation algorithms are not robust enough and often lead to challenging results given surfaces observed under very certain viewing angles. Many of these limitations under such fusion processes can be overcome by casting dense 3D reconstruction and semantic segmentation as a joint optimization formulation, shown in **Fig. 3.** The general idea of the formulation is that each of the voxels gets assigned one out of $L + 1$ labels where label $i = 0$ denotes the free space label and the L



▲ **Figure 2. Overview of the overall architecture for the classical panoptic segmentation (pictures taken from Ref. [26])**



▲**Figure 3. Dense semantic 3D reconstruction**[9]

labels with $i > 0$ indicate the occupied space, which is segmented into several semantic classes. Such formulation, so-called objective function of the volumetric multi-label approaches, can be resolved with the objective function of the convex multi-label energy extended from the volumetric 3D reconstruction energy, as described in Eq. (6). The energy $E(x)$ consists of two parts, in which the former data term is a function of a given label, and is parameterized by the internal probability distribution of the voxel/surfel. The subsequent pairwise smoothness term is a function of the labeling of two connected voxels/surfels in the graph, and is parameterized by the geometry of the map.

$$E(x) = \sum_{s \in \Omega}\left(\sum_{i}\rho_s^i x_s^i + \sum_{i,j:i<j}\Phi_s^{ij}(x_s^{ij} - x_s^{ji})\right), \tag{6}$$

where E represents the objective function of the convex multi-label energy, X_s^i represents the label assigned to voxel s, ρ_s^i represents a cost for assigning label i to voxel s, and Φ_s^i represents transition-specific, direction-and-location-dependent penalizer of the surface area formed as an interface between labels i and j.

This type of formulation describes a convex relaxation procedure, which is closely related to linear programming (LP) relaxations for approximate maximum a posteriori (MAP) estimation inference in Markov random fields (MRFs). The classical solutions to addressing this procedure include the Bayesian, conditional random field (CRF), MRF and variation framework. The work of HÄNE et al.[9] can be referred to for thorough exploration regarding such formations and approaches. HAN et al.[15] also address some latest emerging technique problems, inspired by the continually boosted deep learning achievement.

4 Neural Metric-Semantic Understanding

Following the above overview of the underlying computer vision and machine learning concepts, we will discuss the new explorations regarding fully leveraging the joint semantic & 3D information, neural metric-semantic understanding. Given the high-quality geometric and semantic scene understanding specification, classic semantic and spatial collaborative perception methods can reconstruct global 3D semantic dense maps for a variety of real-world scenes. Moreover, such dense 3D semantic mapping techniques give us explicit editing control over all the elements of the perception pipeline, and strictly follow physical knowledge from computer vision—camera viewpoint, lighting, geometry and materials. However, building high-quality semantic & 3D reconstruction, especially directly from poorly textured areas, under a higher noise level, in dynamic surrounding environments, requires significant manual effort, and automated high consistent context understanding from images is an open research problem. On the other hand, the emerging learning-based techniques are now starting to produce a plausible dense depth map or even 3D scanning of scenes, which is either from random noise or conditioned on certain user specifications. However, they do not yet allow geometrical consistency and cannot always handle the true depth by a single scale factor. In contrast, neural metric-semantic understanding brings the promise of combining these approaches to enable high quality co-consistency under both semantic and geometric scenarios. Neural metric-semantic understanding techniques are diverse, differing in the fusion that they provide over the perception pipeline, the type of fusion and the network structures they utilize. A typical neural metric-semantic understanding approach takes red-green-blue depth (RGBD) sequences corresponding to certain scenes as input, builds a dense 3D reconstruction from them, and adopts the volumetric 3D convolution for point cloud segmentation to extract the final semantic 3D understanding. The dense 3D reconstruction is not restricted by directly using classical computer vision methods to geometric modeling of the environment and can be optimized with the combination of differentiable machine learning techniques for high quality consistent understanding. At the same time, neural metric-semantic understanding approaches incorporate ideas from classical computer vision in the form of orthogonal approaches to reduce drift, traditionally-obtained geometric constraints, and network architectures—to make the learning task easie and the output more consistent.

We propose a taxonomy of neural metric-semantic understanding approaches along the axes that we consider the most important:

- Joint volumetric multi-label formulation
- Semantically geometric and photometric registration
- Semantical depth map regulation

In the following, we will discuss current state-of-the-art methods under these axes.

4.1 Neural Joint Volumetric Multi-Label Formulation

According to the general pipeline of metric-semantic understanding, depth maps and pixel-wise semantic classification scores are achieved as inputs of the final objective, the "semantic understanding of 3D environments". Various approaches are proposed to tackle joint optimization formulation. Authors in Refs. [15, 27 – 28] directly use 3D convolutional neural networks approache on voxels (representation of 3D scenes), like 2D convolution on pixels, while the methods in Ref. [13], such as variational methods for convex relaxation, incorporate the physical knowledge to an emerging differentiable learning network.

3D convolutional neural network methods rely on generic 3D convolutional neural network architectures, and take the three-dimensional representation of 3D scenes as input. The curse of dimensionality applies, in particular, to data that lives on grids, which have three or more dimensions. The number of points on the grid grows exponentially with its dimensionality. In such scenarios, as the counterpart of 2D convolutional processing for two-dimensional pictures, it becomes increasingly important to reduce the computational resources needed for 3D data convolutional processing, such as exploiting sparsity and reduces the number of global memory accesses. Prior work in Ref. [28] implements sparse convolutions (SCs) and introduces a novel convolution operator termed submanifold sparse convolution (SSC) that restricts computation and storage to "active" sites. The utilization of the sparsity nature of points in the 3D volumetric space forms the basis for a new mainstream solution, submanifold sparse convolutional networks (SSCNs), which are optimized for efficient semantic segmentation of 3D representation of scenes. A later trial in Ref. [15] extends the SSCN with explorations in addressing the efficiency bottleneck of sparse 3D CNN, which lies in the unorganized memory access of the sparse convolution steps, for the demand of online computations.

Directly applying 3D convolutional neural networks to voxels like 2D convolution on pixels will introduce some limitations, such as the insufficient capacity of deep learning techniques to delineate visual objects. This, for instance, can result in non-sharp boundaries and blob-like shapes in semantic segmentation tasks. While in the classical perception pipeline, probabilistic graphical models have been developed as effective methods to enhance the accuracy of the above task, as illustrated in Section 3.4. To this end, compared with the classic convex relaxation procedure which always requires regularizers with hand-designed priors, a new differentiable learning network method[13] combines the advantages of classical variational approaches with recent advances in deep learning, and improves the inference/optimization formulation from hand-tuned and not-easy convergence to a simple, generic, and substantially more scalable way. A reason for the improvement is that previously employed priors are not rich enough to capture the complex relationships of our 3D world, while learning-based differentiable networks break through automatically in an end-to-end trainable model. Furthermore, such an explicitly reused concept of variational energy minimization has led to great advances when dealing with noise and missing information.

On a separate track to the progress of joint optimization with neural deep learning techniques, some novel frameworks in Ref. [29] aggregate inputs from the initial stage of the previous pipeline and the information of multiple 2D observations from different view angles, and straightly reconstruct the final 3D semantic results with full deep learning framework. Rather than using the above methods, projecting color data into a volumetric grid and operating solely in 3D, with end-to-end network architecture, directly extracting feature maps from associated RGB images and then mapping into the volumetric feature grid of a 3D network using a differentiable back projection layer can result in more sufficient details.

4.2 Neural Semantically Geometric and Photometric Registration

Despite of the full exploration of the joint optimization formulation with geometry and semantic map as the input, emerging neural network techniques have also tried to leverage the combination of differentiable machine learning techniques with physical knowledge from computer vision in the submodules of the perception pipeline, to enable the classic metric-semantic understanding performance in complex scenes. The seminal methods in Refs.[14] and [31] aim to address the underlying key challenges of such scenarios, namely globally consistent geometric and photometric registration, with some revolutionary thinking, such as fine-tuning the deep network by using the extracted geometric constraints and representing semantics as an invariant scene for medium-term continuous tracking of large scale 3D scanning.

Robust data association is a core problem of visual odometry and the cornerstone of large-scale geometry reconstructions. Currently, the state-of-the-art classic metric-semantic understanding methods use short-term tracking to obtain continuous frame-to-frame constraints, while long-term constraints are established using loop closures, as illustrated in Ref. [14]. Although these two approaches are orthogonal and greatly reduce drift by collaboration, invariant representation of scenes to viewpoint and illumination changes cannot always be guaranteed, because of the gap between action in-

terval spans. The author originally proposes using semantics for medium-term continuous tracking of points to improve the first drift correction strategy. The underlying intuition is that changes in viewpoint, scale, illumination, etc., only affect the low-level appearance of objects but not their semantic meaning. By readily integrating semantic reprojection errors into existing video odometry (VO) approaches and combining differentiable machine learning techniques with physical knowledge from computer vision, translational drift in fast or complex scenes has reduced significantly, as reported in the literature.

The reverse thinking of the above method, emerging as another optimizing direction of deep learning in computer vision, is reflected in the method proposed by LUO et al.[31]. The method leverages a convolutional neural network trained for single-image depth estimation along with conventional structure-from-motion reconstruction to establish geometric constraints on pixels in the image sequence. The authors firstly train a single-image depth estimation network to synthesize plausible depth for general color images, and then fine-tune the network by using the extracted geometric constraints via traditional reconstruction methods at the test time. This novel formula, which combines the strengths of traditional techniques and learning-based techniques, addresses the geometrical consistency of the reconstruction over time even under a gentle amount of dynamic scene motion.

4.3 Neural Semantically Depth Map Regulation

As the basic input of the semantic understanding of 3D environments, input geometry and semantic maps, recorded by the overlapped views or "active" sensing, always suffer from inaccuracy and incompatible resolutions because of the different sensing schemes. Plenty of progress as shown in Refs. [30, 32 – 33] has been made to reduce the noise and boost geometric details, especially after consumer depth sensors coming into our daily lives, marked by the recent integration in the latest iPhone. In many classic metric-semantic understanding approaches, volumetric depth map "fusion" has become a standard method, which shows geometric details boosting with sparse depth and dense RGB information, based on truncated signed distance functions. Due to the disadvantages and the real-time requirement of related classic methods, neural-based novel depth map regulation approaches emerge in multiple ways for new heights of performance: 1) semantic information which enriches the scene representation and is incorporated into the fusion process; 2) leveraging the multi-frame fused geometry and the accompanying high-quality color image through a joint training strategy; 3) depth upsampling method which is tolerant to outlier factors (such as mismeasured depth points, flipping points, and disocclusion) and to spontaneously adapt to each scene by a self-learning framework in an online update manner.

Instead of explicitly combining the geometry and semantic segmentation of 3D in the former, others follow that by including this notion of collaboration more implicitly. However, efficiently encoding and fusing "semantically" rich multi-scale features from an end-to-end trainable (differentiable) way is abnormally obvious. Furthermore, recently there has also been immense progress on learning-based methods that operate on single images. These methods result in the pleasing ability to synthesize plausible depth, in particular, in dynamic scenes as well as limitations of the sensing range. In order to construct fine-grained depth sensing, one of the seminal works by TULSIANI et al[3] specializes those object's representation in scenes to some particular instances, signaling that both top-down and bottom-up cues influence the perception, and perfectly deform into shapes even slightly different from those in the training. Fig. 1 illustrates the pleasing semantic object reconstruction result, which reflects the impressive influence introduced by neural semantic depth map regulation.

5 Applications of Semantic and Spatial Collaborative Perception

Semantic and spatial collaborative perception has many important use cases including, but not limited to, relighting, novel view synthesis, as well as semantic AR contents manipulation. The following is a detailed discussion of various applications.

5.1 Relighting

Relighting is known as a procedure for the photo-realistically rendering of a scene under a novel illumination. It is a fundamental component for a number of media editing applications including AR and visual effects. The previously challenging settings like large-scale outdoor scene relighting can be addressed with the help of multi-view-based semantic and spatial collaborative perception. Relighting in the wild[18] casts the problem as a multi-modal image synthesis problem, which takes a rendered deep buffer as input, containing depth and color channels, together with a semantic label (also known as an "appearance code"), and outputs realistic views of tourist landmarks under various lighting conditions, as shown in **Fig. 4**. Fig. 4a shows that the model is rendered into a deep buffer of depth, color and semantic labels, and Fig. 4b shows that a relighting method translates these buffers into realistic renderings under multiple appearances. The input views including depth and color channels are used to reconstruct the 3D geometry of the scene; the semantic labels are also taken as the input to indicate the location of transient objects like pedestrians. Using the above corresponding rendered deep buffers and pairs of real photos, a multi-modal image synthesis pipeline learns an implicit model of appearance, which represents the time of the day, weather conditions and other properties not presented in the 3D model. A similar principle is also adopted by the multi-view relighting method[34]. Furthermore, the author considers

that such geometry is coarse and erroneous, and directly relighting it would produce poor results. Instead, the geometry is used to construct intermediate buffers—normals, reflection features, and RGB shadow maps—as auxiliary inputs to guide a neural network-based relighting method. The above methods all generalize real scenes, producing high-quality results for applications like the creation of time-lapse effects from multiple images.


(a) Input deep buffer

(b) Output relighting

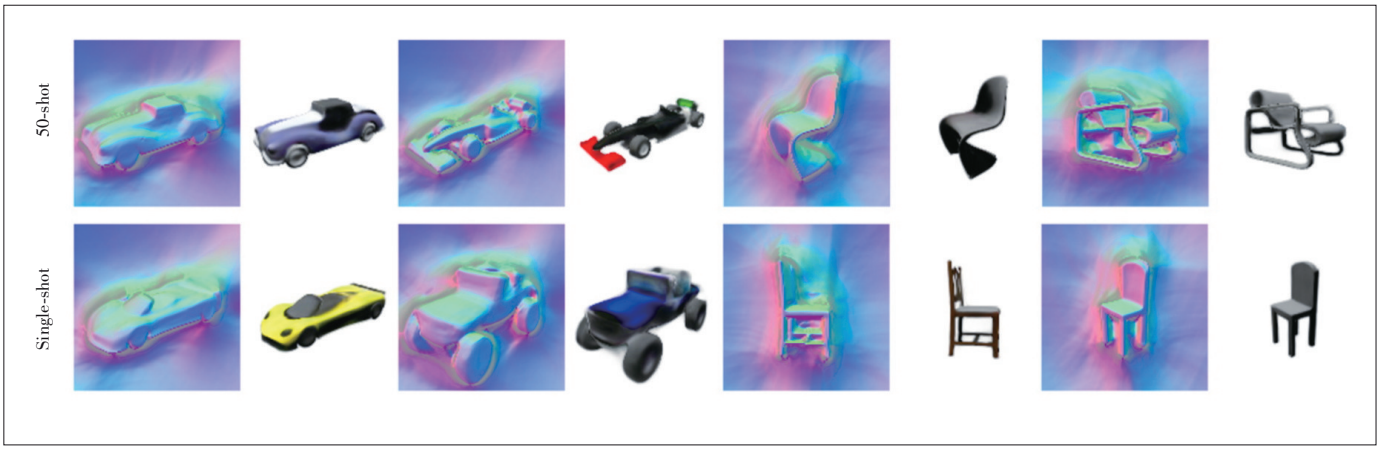
▲Figure 4. Relighting in the wild[18] reconstructs a proxy 3D model from a large-scale Internet photo collection

5.2 Novel View Synthesis

Rendering of a scene under novel camera perspectives of the scene with a fixed set of images given— a procedure known as "novel view synthesis" or "free viewpoint videos"—is a critical component of the emerging media entertainment applications, 360 VR. The topic has gained a lot of interest in the research community and reached compelling quality results with the work of COLLET et al.[35] and its real-time counterpart by DOU et al.[36-37]. Key challenges of such applications are inferring the scene's 3D structure through given sparse observations, for example, the painting of unseen parts of the scene. Recently, reconstructing a learned representation of the scene from the observations, and learning of priors on geometry, appearance and other scene properties in learned feature space with a differentiable renderer, has become a hot topic and made significant progress in previously open challenges such as learning from extremely sparse observations, as shown in **Fig. 5**. Such semantic and spatial collaborative perception-based approaches range from explicit 3D disentanglement of multi-plane images[38] to proposing 3D-structured representations such as voxel grids of features in Refs. [16] and [17]. Among them, HoloGAN[16] implements an explicit affine transformation layer that directly applies view manipulations to learn 3D features to build an unconditional generative model that allows explicit viewpoint changes. Scene representation networks (SRNs)[17] encode both scene geometry and appearance in a single fully connected neural network, to parameterize surface geometry via an implicit function. Although such approaches show better results compared with previous ones, they still have limitations, i.e., they are restricted to a specific use case and limited by the training data.

5.3 Semantic AR Contents Manipulation

Semantic AR contents manipulation is, but not only, the key procedure of the emerging AR experience paradigm, the so-called "retargetable AR"[19]. As the authors illustrate, re-



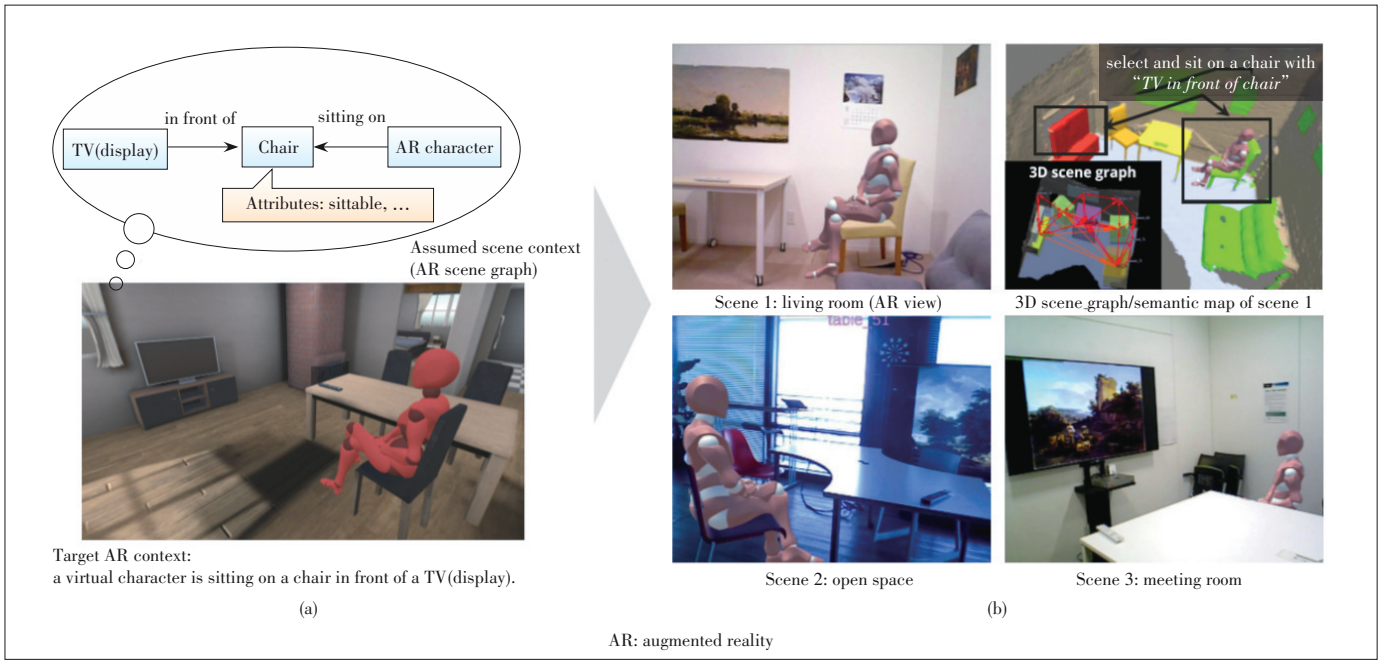
▲Figure 5. Scene representation networks[17] allow full 3D reconstruction from a single image (bottom row, surface normals and color render) by learning strong priors via a continuous, 3D-structure-aware neural scene representation

targetable AR is a novel AR framework that yields an AR experience that is aware of scene contexts set in various real environments, achieving natural interaction between the virtual and real world, as shown in **Fig. 6**, in which images are taken from Ref. [19]. It is expressed as an abstract AR scene graph based on the relationships among objects. Such a retargetable correspondence, which is between the realistic scene and the constructed graph, provides a semantically registered content arrangement, and finally facilitates natural interaction between "digital twins" and the real world. The key procedure, semantic AR contents manipulation, is an extension of the original solution, only a geometric and photometric registration between the virtual and real world[39], to the integration of virtual objects into real environments accurately and naturally. It is achieved by the integration of the advanced abstraction (3D scene graph), and the accurately underlying semantic and spatial collaborative perception, which is the fusion of geometric and semantic information densely reconstructed and labeled in the scene. A similar idea is also proposed by ROSINOL et al.[2], stating that the ideal level of abstraction will be more practical and crucial for the later augmented reality/mixed reality (AR/MR) systems. Even more, linked by such mechanism, the massive knowledge map combined with natural language expressions, and also the above deep understanding of physical environments can be collaboratively learned and managed.

6 Technical Implications

In the above sections, we present a multitude of applications with various target domains by semantic and spatial collaborative perception. While some applications are mostly insensitive to the processing time and response time, others, with legitimate and extremely useful use cases, should be used in an instant reaction manner (e.g., semantic AR contents manipulation). Methods for image and video manipulation are as old as the media themselves, and understanding-based structured visual editing is currently common, for example, in the Internet industry. Neural metric-semantic understanding approaches have the potential to lower the barrier for entry, making manipulation technology accessible to non-experts with limited resources. Although we believe that all the methods discussed in this paper have the potential to positively influence the world via better content creation and storytelling, we must not be complacent. It is important to proactively discuss and devise a plan to systematically arrange the submodules of the above methods under the 5G edge computing scenario for instant reaction and also privacy protection purpose. We believe it is critical that understanding-based synthesizing images and videos are extremely resource- and power-consuming. We also believe that it is essential to raise significant privacy concerns before directly uploading visual raw data to cloud-based semantic and spatial collaborative perception systems, like, for the localization purpose, even if only derived image features are uploaded.

Such related topics regarding "to cloud or not to cloud" were first explored by NAQVI et al.[40], and then extended to edge computing architectures, even with 5G, by BARESI et al.[41–43]. Given the evaluation regarding the added value of cloud computing as a key enabler for AR applications on mo-



▲**Figure 6. Illustration of semantic AR contents manipulation: (a) retargetable AR; (b) framework that retargets the AR scene to various real scenes by comparing the AR scene graph with 3D scene graphs constructed in each of the scenes[19]**

bile devices[40], the authors disclose an important principle that the latency due to connectivity type and the amount of data to be communicated is a major trade-off, and the dynamic deployment and reconfiguration of the framework components between mobile and cloud ends are really important. Furthermore, with respect to the final quality of experience requirements, context-awareness based resource allocation at the wireless network edge[40, 42] and the adoption of serverless edge computing architecture[41 - 43] become the consensus. With the deployment of services to the cloud, the initially widely ignored privacy concerns become an emerging key challenge. The possibility was strikingly demonstrated in Ref. [44], even when only the extracted features are uploaded.

The importance of developing corresponding safe disclosure technologies and building corresponding communities has risen to an urgent position. Such safeguarding measures would reduce the potential for misuse while allowing creative uses of semantic and spatial collaborative perception technologies. In one recent example in the field of image-based localization[45], the authors adopted a cloud-based "obfuscate upload" approach, refraining from uploading the full 3D points of structure-from-motion maps immediately, instead of uploading random line features, lifted from 2D/3D feature points.

Learning from this example, we believe researchers and related business operators must make privacy preserving strategies a key part of all the edge-based semantic and spatial collaborative perception systems with a potential for misuse, but not an afterthought. Also, it is important that we, as a community, continue to develop responsible neural metric-semantic understanding techniques to enable cloud-based semantic and spatial collaborative perception solutions without sacrificing the privacy of users by hiding the privacy concerning contents of the uploading media information.

7 Conclusions

Neural metric-semantic understanding and also the newly neural extension have raised a lot of interest in the past few years. This paper investigates the linkage between the classical and concurrent explorations and a variety of directions related to the topic, which reflects the immense increase of research in this field. The target application is not bound to a specific one but spans a variety of use cases that range from novel view synthesis, relighting, to the manipulation of semantic contents for AR. We believe that metric-semantic understanding will have a profound impact on making complex structured vision understanding and editing tasks accessible to a much broader audience. We hope that this article, which especially focuses on neural metric-semantic understanding, can introduce such modern perception capability to a large research community, which in turn will help to develop the next generation of computer vision applications under the direction.

References

[1] HERMANS A, FLOROS G, LEIBE B. Dense 3D semantic mapping of indoor scenes from RGB-D images [C]//2014 IEEE International Conference on Robotics and Automation (ICRA). Hong Kong, China: IEEE, 2014: 2631 - 2638. DOI: 10.1109/ICRA.2014.6907236

[2] ROSINOL A, ABATE M, CHANG Y, et al. Kimera: an open-source library for real-time metric-semantic localization and mapping [C]//2020 IEEE International Conference on Robotics and Automation (ICRA). Paris, France: IEEE, 2020: 1689 - 1696. DOI: 10.1109/ICRA40945.2020.9196885

[3] TULSIANI S, KAR A, CARREIRA J, et al. Learning category-specific deformable 3D models for object reconstruction [J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(4): 719 - 731. DOI: 10.1109/TPAMI.2016.2574713

[4] TATENO K, TOMBARI F, NAVAB N. Real-time and scalable incremental segmentation on dense SLAM [C]//2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Hamburg, Germany: IEEE, 2015: 4465 - 4472. DOI: 10.1109/IROS.2015.7354011

[5] MCCORMAC J, HANDA A, DAVISON A, et al. SemanticFusion: dense 3D semantic mapping with convolutional neural networks [C]//2017 IEEE International Conference on Robotics and Automation (ICRA). Singapore, Singapore: IEEE, 2017: 4628 - 4635. DOI: 10.1109/ICRA.2017.7989538

[6] NAKAJIMA Y, TATENO K, TOMBARI F, et al. Fast and accurate semantic mapping through geometric-based incremental segmentation [C]//2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Madrid, Spain: IEEE, 2018: 385 - 392. DOI: 10.1109/IROS.2018.8593993

[7] NARITA G, SENO T, ISHIKAWA T, et al. PanopticFusion: online volumetric semantic mapping at the level of stuff and things [C]//2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Macao, China: IEEE, 2019: 4205 - 4212. DOI: 10.1109/IROS40897.2019.8967890

[8] PHAM Q H, HUA B S, NGUYEN T, et al. Real-time progressive 3D semantic segmentation for indoor scenes [C]//2019 IEEE Winter Conference on Applications of Computer Vision (WACV). Waikoloa Village, USA: IEEE, 2019: 1089 - 1098. DOI: 10.1109/WACV.2019.00121

[9] HÄNE C, POLLEFEYS M. An overview of recent progress in volumetric semantic 3D reconstruction [C]//2016 23rd International Conference on Pattern Recognition (ICPR). Cancun, Mexico: IEEE, 2016: 3294 - 3307. DOI: 10.1109/ICPR.2016.7900143

[10] LADICKÝ L, ZEISL B, POLLEFEYS M. Discriminatively trained dense surface normal estimation [C]//European Conference on Computer vision. Zurich, Switzerland: ECCV, 2014: 0906 - 0912. DOI: 10.1007/978-3-319-10602-1_31

[11] GÜNEY F, GEIGER A. Displets: Resolving stereo ambiguities using object knowledge [C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE, 2015: 4165 - 4175. DOI: 10.1109/CVPR.2015.7299044

[12] LI R H, GU D B, LIU Q, et al. Semantic scene mapping with spatio-temporal deep neural network for robotic applications [J]. Cognitive computation, 2018, 10(2): 260 - 271. DOI: 10.1007/s12559-017-9526-9

[13] CHERABIER I, SCHÖNBERGER J L, OSWALD M R, et al. Learning priors for semantic 3D reconstruction [M]//European Conference on Computer vision. Murich, Germany: ECCV, 2018: 325 - 341. . DOI: 10.1007/978-3-030-01258-8_20

[14] LIANOS K N, SCHÖNBERGER J L, POLLEFEYS M, et al. VSO: visual semantic odometry [M]//Computer Vision - ECCV 2018. Cham, switzerland: Springer International Publishing, 2018: 246 - 263. DOI: 10.1007/978-3-030-01225-0_15

[15] HAN L, ZHENG T, ZHU Y H, et al. Live semantic 3D perception for immersive augmented reality [J]. IEEE transactions on visualization and computer graphics, 2020, 26(5): 2012 - 2022. DOI: 10.1109/TVCG.2020.2973477

[16] NGUYEN-PHUOC T, LI C, THEIS L, et al. HoloGAN: unsupervised learning of 3D representations from natural images [C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, South Korea : IEEE, 2019: 7587 - 7596. DOI: 10.1109/ICCV.2019.00768

[17] SITZMANN V, ZOLLHÖFER M, WETZSTEIN G. Scene representation networks: continuous 3D-structure-aware neural scene representations [EB/OL]. [2021-01-05]. https: //arxiv.org/abs/1906.01618

[18] MESHRY M, GOLDMAN D B, KHAMIS S, et al. Neural rerendering in the wild [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recog-

nition (CVPR). Long Beach, CA, USA: IEEE, 2019: 6871 – 6880. DOI: 10.1109/CVPR.2019.00704
[19] TAHARA T, SENO T, NARITA G, et al. Retargetable AR: context-aware augmented reality in indoor scenes based on 3D scene graph [EB/OL]. (2020-08-18) [2021-01-05]. https: //arxiv.org/abs/2008.07817
[20] SCHARSTEIN D, SZELISKI R, ZABIH R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms [C]//Proceedings IEEE Workshop on Stereo and Multi-Baseline Vision (SMBV 2001). Kauai, HI, USA: IEEE, 2001: 131 – 140. DOI: 10.1109/SMBV.2001.988771
[21] NAIR R, RUHL K, LENZEN F, et al. A Survey on time-of-flight stereo fusion [J]. Time-of-flight and depth imaging. sensors, algorithms, and applications, 2013, 8200:105 – 127. DOI: 10.1007/978-3-642-44964-2_6
[22] DAI A, NIEßNER M, ZOLLHÖFER M, et al. BundleFusion [J]. ACM transactions on graphics, 2017, 36(4): 1. DOI: 10.1145/3072959.3126814
[23] WHELAN T, SALAS-MORENO R F, GLOCKER B, et al. ElasticFusion: Real-time dense SLAM and light source estimation [J]. The international journal of robotics research, 2016, 35(14): 1697 – 1716. DOI: 10.1177/0278364916669237
[24] HAN L, FANG L. FlashFusion: real-time globally consistent dense 3D reconstruction using CPU computing [C]//Robotics: Science and Systems XIV. Robotics: Science and Systems Foundation, 2018. DOI: 10.15607/rss.2018.xiv.006
[25] DE GEUS D, MELETIS P, DUBBELMAN G. Fast panoptic segmentation network [J]. IEEE robotics and automation letters, 2020, 5(2): 1742 – 1749. DOI: 10.1109/LRA.2020.2969919
[26] MOHAN R, VALADA A. EfficientPS: efficient panoptic segmentation [EB/OL]. (2020-05-19) [2021-01-05] https: //arxiv.org/abs/2004.02307
[27] ARMENI I, SENER O, ZAMIR A R, et al. 3D semantic parsing of large-scale indoor spaces [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016: 1534 – 1543. DOI: 10.1109/CVPR.2016.170
[28] GRAHAM B, ENGELCKE M, MAATEN L V D. 3D semantic segmentation with submanifold sparse convolutional networks [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018: 9224 – 9232. DOI: 10.1109/CVPR.2018.00961
[29] DAI A, NIENER M. 3DMV: joint 3D-multi-view prediction for 3D semantic scene segmentation [C]//Computer vision. Munich, Germany: ECCV, 2018: 0908 – 0914. DOI: 10.1007/978-3-030-01249-6_28
[30] ROZUMNYI D, CHERABIER I, POLLEFEYS M, et al. Learned semantic multi-sensor depth map fusion [C]//2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). Seoul, South Korea: IEEE, 2019: 2089 – 2099. DOI: 10.1109/ICCVW.2019.00264
[31] LUO X, HUANG J B, SZELISKI R, et al. Consistent video depth estimation [J]. ACM transactions on graphics, 2020, 39(4): 1 – 13. DOI: 10.1145/3386569.3392377
[32] SHIM I, OH T H, KWEON I. High-fidelity depth upsampling using the self-learning framework [J]. Sensors, 2018, 19(1): 81. DOI: 10.3390/s19010081
[33] YAN S, WU C L, WANG L Z, et al. DDRNet: depth map denoising and refinement for consumer depth cameras using cascaded CNNs [C]//European Conference on Computer vision. Murich, Germany: ECCV, 2018. DOI: 10.1007/978-3-030-01249-6_10
[34] PHILIP J, GHARBI M, ZHOU T H, et al. Multi-view relighting using a geometry-aware network [J]. ACM transactions on graphics, 2019, 38(4): 1 – 14. DOI: 10.1145/3306346.3323013
[35] COLLET A, CHUANG M, SWEENEY P, et al. High-quality streamable free-viewpoint video [J]. ACM transactions on graphics, 2015, 34(4): 1 – 13. DOI: 10.1145/2766945
[36] DOU M, KHAMIS S, DEGTYAREV Y, et al. Fusion4D: Real-time performance capture of challenging scenes [J]. ACM transactions on graphics, 2016, 35(4): 1 – 13. DOI: 10.1145/2897824.2925969
[37] DOU M S, DAVIDSON P, FANELLO S R, et al. Motion2Fusion [J]. ACM transactions on graphics, 2017, 36(6): 1 – 16. DOI: 10.1145/3130800.3130801
[38] XU Z X, BI S, SUNKAVALLI K, et al. Deep view synthesis from sparse photometric images [J]. ACM transactions on graphics, 2019, 38(4): 1 – 13. DOI: 10.1145/3306346.3323007
[39] KIM K, BILLINGHURST M, BRUDER G, et al. Revisiting trends in augmented reality research: a review of the 2nd decade of ISMAR (2008 – 2017) [J]. IEEE transactions on visualization and computer graphics, 2018, 24(11): 2947 – 2962. DOI: 10.1109/TVCG.2018.2868591
[40] NAQVI N Z, MOENS K, RAMAKRISHNAN A, et al. To cloud or not to cloud: a context-aware deployment perspective of augmented reality mobile applications [C]//Proceedings of the 30th Annual ACM Symposium on Applied Computing. Salamanca Spain. New York, USA: ACM, 2015: 0413 – 0417. DOI: 10.1145/2695664.2695880
[41] BARESI L, FILGUEIRA MENDONÇA D, GARRIGA M. Empowering low-latency applications through a serverless edge computing architecture [C]//Service-oriented and cloud computing. Oslo, Norway: ESOCC, 2017: 0927 – 0929. DOI: 10.1007/978-3-319-67262-5_15
[42] CHATZIELEFTHERIOU L E, IOSIFIDIS G, KOUTSOPOULOS I, et al. Towards resource-efficient wireless edge analytics for mobile augmented reality applications [C]//2018 15th International Symposium on Wireless Communication Systems (ISWCS). Lisbon, Portugal: IEEE, 2018: 1 – 5. DOI: 10.1109/ISWCS.2018.8491206
[43] BARESI L, FILGUEIRA MENDONÇA D. Towards a serverless platform for edge computing [C]//2019 IEEE International Conference on Fog Computing (ICFC). Prague, Czech Republic: IEEE, 2019: 1 – 10. DOI: 10.1109/ICFC.2019.00008
[44] PITTALUGA F, KOPPAL S J, KANG S B, et al. Revealing scenes by inverting structure from motion reconstructions [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019: 145 – 154. DOI: 10.1109/CVPR.2019.00023
[45] GEPPERT M, LARSSON V, SPECIALE P, et al. Privacy preserving structure-from-motion [C]//16th European Conference Computer Vision. Glasgow, United Kingdom: EVVC, 2020:0823 – 0828. DOI: 10.1007/978-3-030-58452-8_20

Biography

ZHU Fang (zhu.fang@zte.com.cn) received the B.Eng. degree in electrical engineering and the M.Sc. degree in information and system from Xi'an Jiaotong University, China and the Ph.D. degree in electronic engineering from Southeast University, China. He is currently the director of the technical committee in digital video and vision of ZTE Corporation, and also the deputy director in multimedia of State Key Laboratory of Mobile Network and Mobile Multimedia Technology. He is a senior member of IEEE, focusing on circuits and systems for video technology and smart vision. His research interests include core technology, cloud architecture and acceleration chipset for specific application of XR & Smart Vision based on mobile computing.