Normal Assisted Pixel-Visibility Learning With Cost Aggregation for Multiview Stereo

Wei Tong, Xiaorong Guan, Jian Kang, *Member, IEEE*, Poly Z. H. Sun, *Member, IEEE*, Rob Law, Pedram Ghamisi, *Senior Member, IEEE*, and Edmond Q. Wu, *Senior Member, IEEE*

***Abstract*— Multiple-View Stereo (MVS) aims to reconstruct the dense 3D representations of scenes. MVS has potential applications in the fields of autonomous driving (unstructured environment construction) and robotic navigation (visual-inertial navigation). To mitigate the error of depth estimation in low-textured or occluded regions, this work proposes a two-stage multi-view stereo network for fast and accurate depth estimation. The improvements of this work over the state of the art are as follows: 1) Sparse costs are constructed to jointly predict the initial depth map and surface normal by cost regularization, which proves that the surface normals can be estimated in this way with low memory consumption. 2) A new edge refinement block is developed to refine the coarse surface normal to obtain a fine-grained surface normal map. 3) Instead of using the general variance-based metric to equally aggregate cost, a new content-adaptive cost aggregation mechanism based on the similarity of the neighboring surface normal is designed for reliable cost aggregation. To the best of our knowledge, the proposed work is the first trainable network that leverages surface normal as guidance to capture neighboring pixel-visibility, which is an effective supplement to existing depth/normal estimation frameworks. Experimental results indicate that our method can not only achieve accurate depth estimation for scene perception but also make no concession to the real-time performance and limited memory bottleblock. Multiple-view stereo (MVS) aims to reconstruct the dense 3D representations of scenes. It is widely used in the fields of industrial measurement, autonomous driving, and robotic navigation. To mitigate the error of depth estimation in challenging scenarios, this work proposes a two-stage multi-view stereo network for fast and accurate depth estimation. Our method is the first trainable network that leverages surface normal as pixel-visibility guidance to aggregate reliable cost, which could achieve accurate depth estimation and provide the perception ability for the robot. The proposed method has great potential in the fields of 3D reconstruction, industrial measurement, and robotic navigation to estimate real-time and accurate depth with limited memory consumption.**

***Index Terms*— Cost aggregation, depth estimation, multi-view stereo, pixel visibility, surface normal.**

Manuscript received 30 January 2022; revised 14 April 2022 and 25 June 2022; accepted 20 July 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 72192820, Grant 72192824, Grant 62171274, and Grant U1933125. The Associate Editor for this article was L. M. Bergasa. *(Corresponding authors: Xiaorong Guan; Jian Kang; Edmond Q. Wu.)*

Wei Tong and Xiaorong Guan are with the School of Mechanical Engineering, Nanjing University of Science and Technology, Jiangsu, Nanjing 210094, China (e-mail: tongwei@njust.edu.cn; gxr@njust.edu.cn).

Jian Kang is with the School of Electronic and Information Engineering, Soochow University, Suzhou 215006, China (e-mail: kangjian_1991@outlook.com).

Poly Z. H. Sun is with the Department of Industrial Engineering, School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China, and also with the Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: zh.sun@sjtu.edu.cn).

Rob Law is with the Asia-Pacific Academy of Economics and Management, University of Macau, Macau 999078, China, and also with the Department of Integrated Resort and Tourism Management, Faculty of Business Administration, University of Macau, Macau 999078, China (e-mail: roblaw@um.edu.mo).

Pedram Ghamisi is with the Helmholtz-Zentrum Dresden-Rossendorf (HZDR), Helmholtz Institute Freiberg for Resource Technology (HIF), Exploration, D09599 Freiberg, Germany, and also with the Institute of Advanced Research in Artificial Intelligence (IARAI), 1030 Vienna, Austria (e-mail: p.ghamisi@gmail.com).

Edmond Q. Wu is with the Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, China, and also with the Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai 200240, China (e-mail: edmondqwu@163.com).

Digital Object Identifier 10.1109/TITS.2022.3193421

I. Introduction

INTELLIGENT vehicle is an important part of intelligent transportation system, the specific unmanned environment sensing technology has attracted extensive attention in both academia and industry [1]. As a scarce resource in this field, real-time positioning of intelligent vehicles and dense high-precision environment reconstruction technology plays a core role in the whole field, which can help unmanned vehicles perceive the complex information of the road environment in advance [2], [3]. Among them, Multiple-View Stereo (MVS) aims to infer accurate scene depth information and realize dense scene reconstruction from captured adjacent images. Previous studies have embedded MVS into real-time Simultaneous Localization and Mapping (SLAM) in the application of intelligent transportation systems. According to the self-motion estimation of SLAM and depth estimation of MVS, more accurate and dense scene information can be perceived, which has great practical application value than the original SLAM. In addition, assisted depth information has also been widely used in the navigation system and measurement tasks. Navigation methods [8], [42] have been proposed to perceive the environment and localize the robot by assisted monocular and binocular vision. Li and Wang [19] presented a robust stereo method to measure the 3D shape of dynamic objects. Cai *et al.* [48] proposed a pose-only representation for the multiple-view imaging geometry, which can further improve the accuracy and robustness of visual reconstruction and navigation. In [25], a new radiation-free robotic ultrasound system was developed to measure the spine under assisted monocular vision. Compared with Structure From Motion (SFM) [23], MVS methods can generate dense 3D point clouds. Its specific steps are as follows: 1) given pre-calibrated camera parameters beforehand using SFM, the

depth values of each view are discretized and searched to realize the highest evaluation of photometric consistency among neighboring images. 2) depth values are back-projected to the reference view to obtain its corresponding 3D point cloud. 3) this will be fused to obtain final point clouds with camera parameters.

Traditional MVS methods only consider the photometric consistency without adding other visual cues such as lighting, shadows, or semantic information. Benefiting from the powerful feature extraction ability of deep learning, learning-based methods [4], [5] extracted local and global features from multiple RGB images to directly infer the depth map. Particularly, Yao *et al.* proposed MVSNet [9] to extract discriminative image features, and then implicitly encoded the local prior geometry such as specular reflection to construct 3D cost volume, which is further regularized to estimate final depth. Its follow-up methods [10], [13], [29] have realized significant improvement on MVS bench-marks [4], [6], [7] over especially traditional MVS methods.

However, there is still a challenging task to be tackled for current learning-based MVS methods, where the task is undertaken within conditions in uncontrolled environments, such as illumination, specular reflectance, scene variability, and incorrectly registered views. Therefore, the wrong visibility for cost aggregation may inevitably deteriorate the final reconstruction. To tackle these uncontrolled factors, DeepMVS [37] proposes a patch aggregation network to calculate image disparity values. FADE [15] improves stereo matching accuracy by introducing the fusion of multi-level aggregation in spatial and context information, respectively. In addition, [16], [31] additionally apply the max pooling and averaging operation to generate the pixel-visibility map for multi-scale cost volume aggregation. Nevertheless, these types of methods based on adaptive cost aggregation can be time-consuming in constructing patch plane volumes and require additionally designed modules to learn the pixel-wise visibility required for cost aggregation.

Given the above problems, a multi-stage multi-view stereo network is proposed in this paper. Inspired by GeoNet++ [17], which jointly predicts depth map and surface normal map by 3D consistency constraint, we consider that surface normal can intuitively express local geometry attributes to some extent, which is insensitive even in low-textured regions. Therefore, surface normal can be easily estimated from visual appearance compared with depth estimation. Different from previous works [20], [21] that separately predict depth and surface normal by two networks, which potentially results in inconsistent predictions and increases the complexity of the model. As a compromise, we design a new two-stage network architecture to realize more robust cost aggregation by local surface normal.

Our main contributions are as follows: 1) A two-stage learning-based stereo network based on pixel-visibility is proposed to jointly estimate a depth map and surface normal. 2) A new cost aggregation mechanism is proposed to aggregate reliable costs for depth inference. This is the first work to learn pixel-visibility through the similarity of the neighboring surface normal. 3) The experiment shows that surface normal can be effectively estimated by multi-view cost volume regularization. 4) Given sparse cost volume and limited memory bottlenecks, our work achieves comparable performance with existing state-of-the-art works on the public DTU and Tanks and Temples datasets.

II. Related Work

A. Traditional MVS

Traditional MVS methods can be divided into the following categories: depth map-based [22]–[24], voxel-based [41], surface-based [26], and patch-based [27]. Compared with other methods, depth map-based methods perform more flexibly and concisely and mainly focus on exploiting various representations from low-level image cues to geometric constraints. Specifically, these methods include matching cost calculation, aggregation, and depth value regression. For example, Saxena *et al.* [18] leveraged Markov random fields to predict depth maps by hand-crafted features from a single image. Schönberger and Frahm [23] proposed COLMAP to jointly estimate depth map and surface normal by selecting neighboring views. Although this method has become the standard of the traditional MVS methods due to its high accuracy on the benchmarks, its entire implementation process is time-consuming. In addition, although all of these traditional methods can work well in relatively simple scenarios, they are difficult to adapt to other scenarios.

B. Learning-Based MVS

Recently, deep learning has achieved great success in the field of computer vision, brain-computer interface and natural language processing, such as computer-aided diagnosis [44], [46], [47], semantic segmentation [43], [45], text processing [49] and brain cognitive science [50], [51], [52]. More learning-based convolutional neural network (CNN) studies have tended to exploit MVS without using traditional hand-crafted image features. To deal with more complex lighting like textureless and occluded scenes, depth map-based methods apply CNNs to extract pixel-wise features instead of using the hand-crafted features; this method yields more robustness than the traditional MVS. Wang *et al.* [14] proposed an efficient multi-view stereo framework based on the traditional patch-match algorithm. For the global method, MVSNet [9] is a representative network for accurate multi-view depth inference. It implicitly encodes multi-view camera geometries through the use of differential homography warping to construct cost volume. Then, 3D convolutions are used to regularize cost volume to measure the multi-view similarity. Follow-up works, such as CasMVSNet [29] and UCSNet [30], construct cost volume in a coarse-to-fine manner to ease the memory limitation. To prevent the problem of matching ambiguity, PVA-MVSNet [16] and PVSNet [31] apply the pixel-wise visibility mechanism to aggregate reliable matching costs. However, it is not easy to explicitly implement geometric constraints in a designed network [32]. These methods can easily lead to wrong visibility for cost aggregation, especially in the illumination and occluded areas, and the quality of the final point cloud reconstruction inevitably deteriorates.

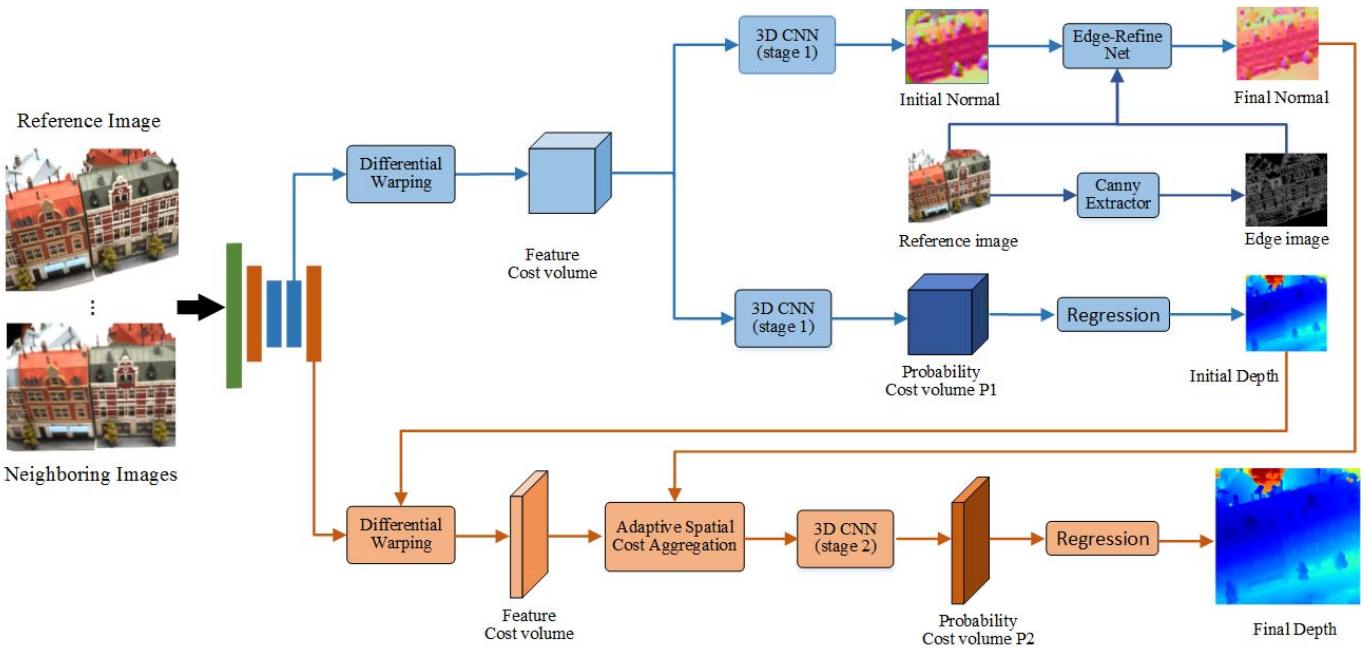


Fig. 1. The proposed two-stage multi-view stereo network architecture. A coarse depth map and surface normal are obtained in the first stage(shown in blue). Accurate depth estimation is obtained in the second stage (shown in orange) with the assisted surface normal for cost aggregation.

C. Surface Normal Estimation

Surface normal represents an important geometry property for understanding a 3D scene. Recently, several data-driven methods have learned image features and textures to predict the normal of a single image and have achieved promising results. Wang *et al.* [21] propose a surface normal prediction network with the incorporation of local, global, and vanishing point information. Zeng *et al.* [33] develop a hierarchical fusion network with adaptive feature reweighting to predict the single image surface. To simplify the network architecture in our study, the surface normal of the reference image is directly estimated by the multi-view cost regularization module, rather than designing a branch separately as in previous works. Then surface normal is used as the position-specific weight for reliable cost aggregation to realize depth inference.

III. Methods

To reconstruct a complete and accurate point cloud with fine-grained 3D geometry for MVS, a two-stage network for fast and accurate depth inference is proposed in this work. The overall network architecture is illustrated in Fig.1, and is divided into two modules. The first module (shown in blue) jointly predicts initial depth value and surface normal from sparse cost volume, which is aggregated by a general variance-based metric. The second module (shown in orange) leverages the refined normal map with the neighboring pixels' guidance to aggregate reliable cost volume for accurate depth inference. Each component of this work is described in the following sections.

A. Multi-Level Feature Extraction

Supervised learning-based MVS methods assume that the input images are already pre-calibrated and the network input is composed of a reference image and its adjacent source images. Previous works mainly extract high-level semantic feature maps to construct standard cost volume by homography warping but lack the fine representations of the low level. Therefore, the 2D UNet feature extractor is applied to obtain a finer representation via encoding and decoding operations. As shown in Fig. 1, for efficient computation, the size of the estimated depth maps in the two stages are $\{1/4, 1/2\}$ of the reference image, and the channels of the corresponding feature maps are 32 and 16, respectively. To ensure reasonable high-frequency feature extraction in two-stage depth estimation, the learned upsampling operation is applied to properly incorporate low-resolution information with high-resolution features. Since 3D convolution is usually time-consuming and requires a large quantity of GPU memory, the sparse cost volume is constructed on the corresponding feature maps to predict the depth map. In addition, the 3D regularization module is used to predict surface normal only in the first stage.

B. Joint Estimation Between Depth and Surface Normal

Recent works on surface normal prediction mainly focus on single-view scenes. It is not easy to directly predict accurate and consistent surface normal geometry by a monocular feature extraction network. In the following section, we start a discussion with plane sweep volume construction and cost regularization, and then describe the detailed implementation of depth inference and surface normal estimation with multiple views.

1) Plane Sweep Volume Construction: To construct a multi-view cost volume for $1/4$ scale of the reference image, the feature maps of the source view are warped to the reference view by camera coordinate conversion. This transformation is performed through front-parallel planes with multiple sampling depth hypotheses. The coordinate relationship obtained

by homography mapping is as follows:

$$M_i(d) = K_i \cdot R_i \cdot (I - \frac{(t_1 - t_i) \cdot n_1^\top}{d}) \cdot R_1^\top \cdot K_1^{-1}, \quad (1)$$

where $M_i(d)$ represents the homography matrix conversion from the source feature maps i^{th} to the reference view with the uniformly sampling depth hypothesis interval $[d_{min}, d_{max}]$. In addition, n_1 refers to the reference camera principal axis, K_i, R_i, t_i represents intrinsic parameters, relative rotation matrix, and relative translation matrix, respectively, for the i^{th} view. The cost volume construction in the first stage is similar to the previous works [10], [12], [34], which equally treat all pixels and apply a variance-based metric to aggregate standard plane sweep volume. This approach can support the arbitrary number of multiple views. The feature volume $\{V_i\}_{i=1}^N$ of all-view is aggregated to a cost C_{agg}. Its size is $\frac{W}{4} \times \frac{H}{4} \times D \times C$ in the first stage, where W, H, D, and C represent the width, height, sampling depth number, and the channel of feature maps, respectively.

2) Depth Estimation Based on Cost Volume: Note that the aggregated cost inevitably contains noise due to the influence of a non-Lambertian surface [9]. To solve this problem, multi-scale 3D CNNs are utilized to infer the corresponding relationship of feature matching and predict the probability distribution of depth value. Considering its effectiveness, 3D UNet is adopted for cost regularization, as it can efficiently infer scene geometry on multiple scales and learn the importance of different depth hypotheses.

After generating the probability volume $P = \frac{W}{4} \times \frac{H}{4} \times D$ by the cost regularization, the $soft\ argmin$ operation is selected to estimate the depth value, and is defined as follows:

$$D_{est} = \sum_{d=d_{min}}^{d_{max}} d \cdot P(d). \quad (2)$$

3) Surface Normal Estimation Based on Cost Volume: Inspired by the work in Geonet++ [17], the estimation of the surface normal $n = [n_x, n_y, n_z]$ for the pixel i can be regarded as a least-square problem. Therefore, the well-trained neural network can theoretically predict the corresponding surface normal. In addition, depth estimation and surface normal are mutually restricted. The depth value is constrained by the local tangent plane of the surface normal, and the actual surface normal can be directly computed by its neighboring depth value. Therefore, a network is designed to decode the geometric properties for depth estimation and surface normal. Since the standard plane sweep volume based on uniformly sampling depth hypotheses contains rich semantic and geometric information, the sparse plane scanning volume with a size of $\frac{W}{4} \times \frac{H}{4} \times D \times C$ is used to predict not only the depth but also the coarse surface normal in the initial stage. Note that a regularization module similar to that in-depth estimation is applied to regularize the cost volume for surface normal estimation. In the process of regularization, the cost is squeezed along the direction of sampled depth rather than the channel direction, and the size of regularized cost is $\frac{W}{4} \times \frac{H}{4} \times C$. This operation can be regarded as an encoder that squeezes potential spatial geometric characteristics along sampled depth and simultaneously suppresses noise. Then three 32-channel 2D dilated convolution layers with a kernel size of 3×3 are used, and three ordinary convolution layers follow. The size of the surface normal predicted in the initial stage by the decoding operation of these convolution layers is $\frac{W}{4} \times \frac{H}{4} \times 3$.

C. Surface Normal Edge Refinement Module

Since the sparse cost volume contains more candidates along the depth direction and has no significant features in the non-textured regions, the initial surface normal still needs to be further optimized. Note that the cost aggregation mechanism for depth estimation is proposed in Section III.D, and is based on the predicted surface normal to improve the accuracy of depth estimation. The reference image and the edge image contain a large number of visual geometry cues. To suppress the noise and make the boundary prediction more accurate for surface normal estimation, we use the reference image and the edge image extracted by the canny operator as guidance to refine the surface normal. As illustrated in Fig. 2, the reference image, the edge image, and the initially predicted surface normal are concatenated as 9-channel inputs. The inputs are passed through seven 2D dilated convolution layers, and the receptive field increases continuously as the number of layers deepens. The refined surface normal is used for the cost volume aggregation mechanism in the second stage.

D. Adaptive Cost Volume Aggregation Based on Surface Normal

The depth map estimated in the first stage may be incorrect because the sparse cost volume has some incorrect matching correspondence. Therefore, we use the initial depth map to construct a more targeted plane sweep volume in the second stage. The sampling depth range and depth hypothesis become narrower than in the first stage, which can support high-resolution cost construction under a limited memory bottleneck. We believe that it is unreasonable to equally treat the contribution of all pixels for cost aggregation by using variance-based metrics, especially in the occluded and illumination regions. Furthermore, the neighboring pixels are usually highly correlated and depth values are prone to be continuous, while depth values usually vary drastically on the borders of the object. To further improve the precision of depth estimation in the non-textured and edge-fattening regions, we leverage surface normal to design a cost volume aggregation mechanism.

Previous works either assumed that all pixels contributed equally to cost aggregation, or designed additional branches to learn the pixel-wise visibility for reliable cost aggregation. The proposed adaptive cost aggregation mechanism is determined by surface normal, which is simpler. The similarity measurement mechanism of the neighboring surface normal for cost aggregation is illustrated in Fig. 3, where $n_i = [n_{ix}, n_{iy}, n_{iz}]$ represents the normal value of a current pixel in the surface normal map. The cosine similarity is selected as a linear kernel to characterize the local geometric characteristics of eight neighbors. In particular, if the surface normal angle between

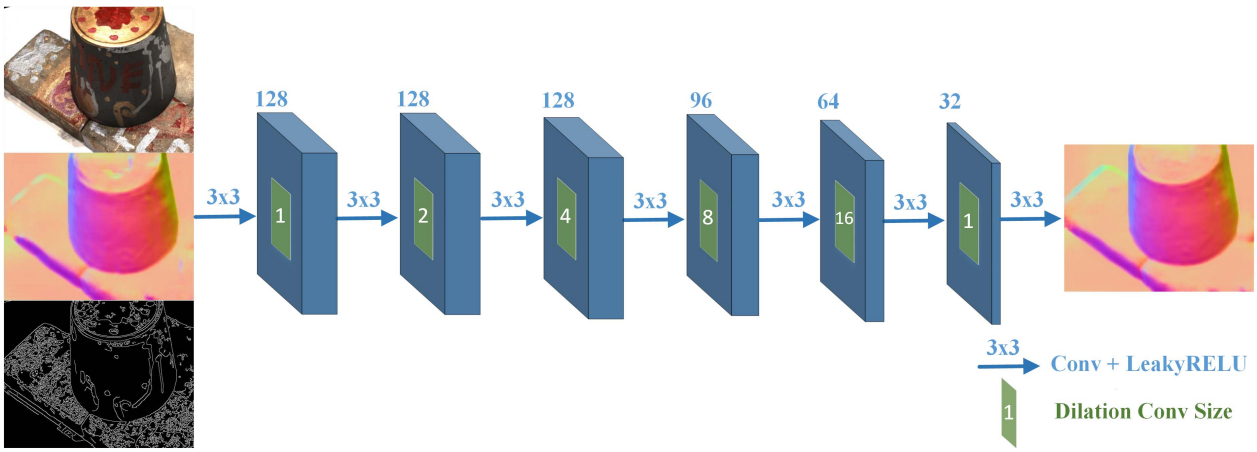


Fig. 2. The details of the surface normal edge refinement module.

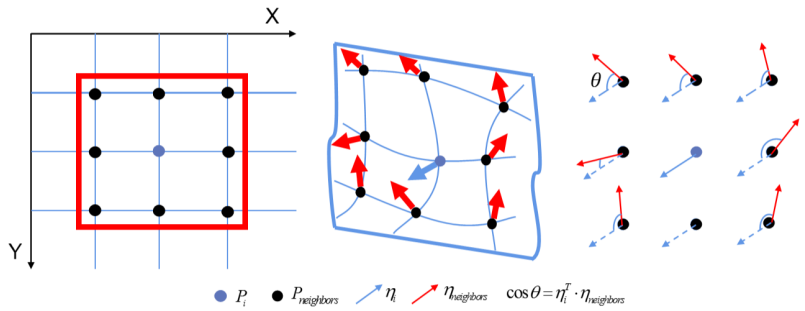


Fig. 3. The proposed similarity measurement mechanism of the neighboring surface normal for cost aggregation.

the current pixel n_i and its neighboring pixel n_j is small, it indicates that the two pixels are located on the same tangent plane with a high probability. As a result, the pixel j makes more of a contribution to the depth estimation of the pixel i on the weight map of cost aggregation.

For the pixel-visibility weight map to aggregate spatial cost over multiple cost volumes, eight neighbors are considered. To suppress incorrect matching correspondence, especially on some edge or irregular surface, the weight w_j, based on the relationship of the neighboring surface normal, is introduced to constrain the geometric consistency of the pixel visibility. The weight w_j is defined as:

$$w_j = e^{-\alpha_1 |n_i^\top n_j|}. \tag{3}$$

The weight w_j depends on the cosine similarity between its neighboring pixels. A high cosine similarity value indicates that the pixel-visibility and depth value of the pixel j is more related to the pixel i. Each two-view cost is first generated by the weighted sum of eight neighbors and summed with the residual of the original two-view cost. Compared with AANet [35], which designed an independent network to learn the pixel-visibility relationship for non-textured regions, our cost aggregation mechanism based on surface normal is not strict for the accuracy of the surface normal and has stronger robustness and noise resistance. Since each two-view cost is defined by the aggregation operation in the second stage, we only average the cost of all views, rather than using the variance-based metric. The final cost volume C'_{agg} is defined as:

$$w'_j = \frac{w_j}{\sum_{j=1}^{8} w_j}, \tag{4}$$

$$C'_{agg} = \sum_{i=1}^{N-1} \frac{1}{N-1} \left(\sum_{j=1}^{8} w'_{i,j} C_{neighbor}\right) + C_{ref}. \tag{5}$$

E. Multi-Metric Loss

In the following, we introduce the loss function used for network training. The initial surface normal, the refined surface normal, and the ground-truth in the first stage are specified as n_i, n'_i, and n_i^{gt}, respectively. Similarly, d_{i1}, d_{i2}, and d^{gt} represent the estimated depth value in the first stage, the estimated depth value in the second stage, and the corresponding ground-truth, respectively. The overall loss function is expressed as $L_{sum} = L_{stage1} + L_{stage2}$, which is composed of the depth loss and surface normal loss. Both depth loss and surface normal loss are measured based on the mean absolute difference. The loss in the first stage is expressed as:

$$L_{stage1} = \frac{\lambda_1}{M}\Big(\sum_i ||n_i - n_i^{gt}||_1 + \sum_i ||n'_i - n_i^{gt}||_1 + \sum_i ||d_{i1} - d^{gt}||_1\Big), \tag{6}$$

where M denotes the mask of valid pixels.

The loss in the second stage is defined as:

$$L_{stage2} = \frac{\lambda_2}{M}(\sum_i ||d_{i2} - d^{gt}||_1). \quad (7)$$

Note that λ_1 and λ_2 are hyper-parameters to adjust the loss contribution of the two stages, losses in network training. λ_1 is set to 1, and λ_2 is set to 2, which are not sensitive to the performance of model training.

IV. Experiment Results

In this section, we evaluate the effectiveness of the proposed method with a detailed ablation study on the indoor DTU datasets [36] and the outdoor Tanks and Temples datasets [7].

A. Datasets and Accuracy Metrics

The large-scale MVS datasets called DTU include 124 different scenes, which are scanned under 7 different lighting conditions with given camera parameters. Every scan has 49 views, and its image resolution is 1600×1200. To fairly compare with other methods, we follow the data partition manner used [9], [26] to choose the same training, validation, and evaluation sets. The screened Poisson surface reconstruction is used to generate the ground truth depth maps for supervised training. The Tanks and Temples datasets are selected to validate the generalization ability of the trained model trained. The Tanks and Temples datasets are collected from real video sequences with a small depth range, and their image resolution is resized to 1920×1056. In addition, it includes both intermediate and advanced datasets. In this work, we select the intermediate sets containing 8 scenes for performance evaluation.

Accuracy metrics: To quantitatively evaluate the performance of the reconstructed point clouds on DTU datasets, we choose the standard metrics, including the mean completeness (named Comp.), the mean accuracy (named Acc.), and the overall accuracy is defined as:

$$Overall = \frac{1}{2}(Acc. + Comp.). \quad (8)$$

The specific accuracy and completeness can be calculated by the official Matlab code provided by DTU datasets. In addition, the F-score metric is chosen to evaluate the generalization ability on Tanks and Temples datasets, which is defined as:

$$F(d) = \frac{2 * P(d)R(d)}{P(d) + R(d)}, \quad (9)$$

where $P(d)$ is the accuracy under a given distance threshold d, and $R(d)$ is the recall of point cloud reconstruction. F-score is the harmonic average of accuracy and recall.

B. Implementation Details and Post Processing

The proposed method is trained on the DTU datasets and implemented by the framework of PyTorch. The number of source views is set to 4 and the input image resolution is 640×512; the same view selection strategy as in MVSNet is used in our work. The number of depth hypotheses for the two stages are 48 and 32, respectively. The resolution of the depth maps predicted in the two stages are 1/4 and 1/2 of the reference image, respectively. The Adam optimizer (with $\beta_1 = 0.9$, $\beta_2 = 0.999$) is selected to train the network for 16 epochs. The initial learning rate is set to 0.001 and will be halved at the 8^{th}, 10^{th}, and 12^{th} epochs to prevent the learning parameters from falling into local optima. The batch size is set to 4 on a device with NVIDIA GTX 3090 GPU.

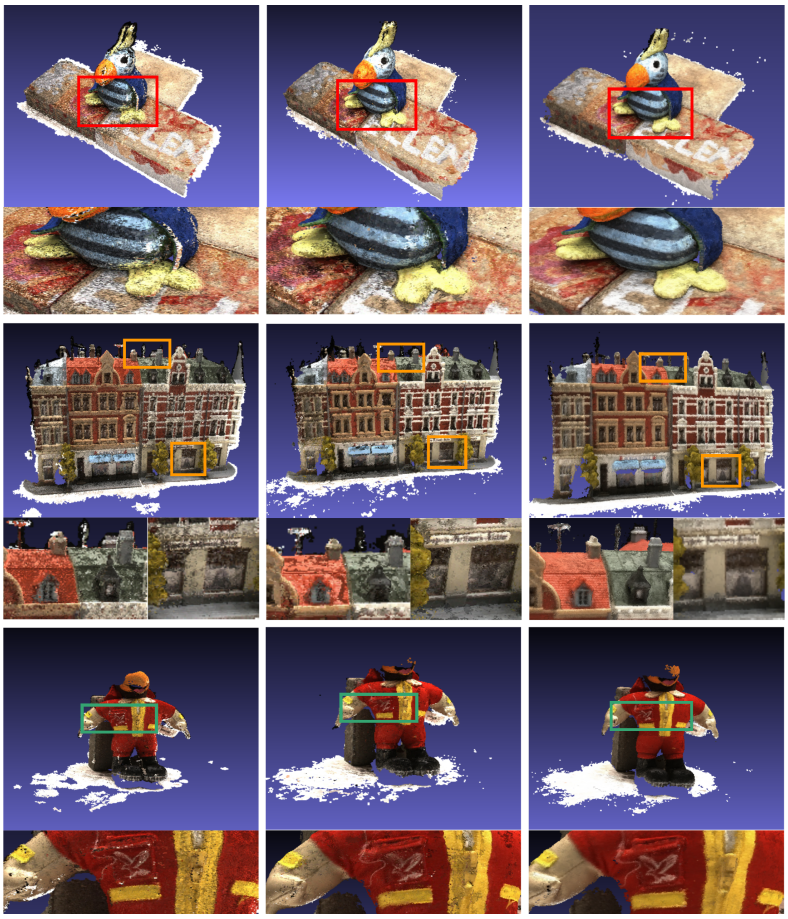
Post-processing: After obtaining the predicted depth maps, we followed [9] to fuse the depth maps of all views to generate dense point clouds. It includes three steps: photometric filtering for outlier elimination, geometric consistency for depth consistency constraint, and depth maps of different views are fused into the final point cloud representation.

TABLE I
Quantitative Comparisons of Different Methods on DTU Evaluation Sets

Methods	Comp- (mm)	Acc- (mm)	Overall	GPU Mem (MB)	Run time (s)
Camp [6]	0.554	0.835	0.695	-	-
Tola [32]	1.190	0.342	0.766	-	-
Gipuma [22]	0.873	**0.283**	0.578	-	-
SurfaceNet [26]	1.040	0.450	0.745	-	-
MVSNet [9]	0.646	0.456	0.551	10823	1.21
R-MVSNet [10]	0.452	0.383	0.417	7577	1.28
MVSCRF [39]	0.426	0.371	0.398	5430	1.8
Point-MVSNet [13]	0.421	0.361	0.391	8731	3.35
Fast-MVSNet [11]	0.403	0.336	**0.370**	5300	0.6
CasMVSNet [29]	0.406	0.396	0.401	4093	0.243
Our 1st stage	0.586	0.612	0.599	**2821**	**0.135**
Our method	**0.367**	0.375	0.371	4655	0.344

C. Benchmark Performance

1) Evaluation on DTU Datasets: We conduct comprehensive experimental comparisons on the DTU testing set, which contains 22 scenes. The average value of accuracy and completeness on all testing sets are selected as specific metrics to evaluate the overall point cloud reconstruction quality. The sampling depth hypotheses are limited to between 425mm and 935mm. In particular, given the same input, the size of the depth map predicted by Point-MVSNet [13] and Fast-MVSNet [11] is consistent with the original resolution, while the depth map predicted by MVSNet and R-MVSNet is only 1/4 of the original resolution. To realize fast and accurate depth estimation with a simple network, we only construct cost volume at 1/4 and 1/2 resolution to infer the two-stage depth map. Table I shows the quantitative comparison results of different methods. It can be seen that Gipuma [22] performs well in terms of accuracy metrics. Although the depth map predicted by our method is only half of the Fast-MVSNet [11], our method achieves the best performance with respect to overall accuracy and completeness. In particular, our depth estimation network follows the two-stage Cas-MVSNet [29] and the depth value sampling parameters are similar. It can be seen that the memory consumption and running time are slightly higher than for Cas-MVSNet, due to the introduction of surface normal estimation and cost aggregation mechanism. However, the overall reconstruction performance of our method



a) R-MVSNet [7] b) Ours without AM c) Ours with AM

Fig. 4. Qualitative comparisons of point cloud reconstruction on DTU datasets.

is better than the two-stage Cas-MVSNet, which validates the effectiveness of our proposed normal-assisted depth estimation module. Qualitative comparison with R-MVSNet is shown in Fig. 4. It can be seen that through the proposed aggregation mechanism, our model can generate a more complete point cloud with finer and cleaner details, especially in the object boundary and texture-less regions (*e.g.,* top edge and door-frame in the second row of Fig. 4). This might be attributed to the anti-noise characteristics of surface normal, which can construct reliable cost through the weighted aggregation of eight neighbors.

2) Memory and Run-Time Comparison: We continue to evaluate the proposed method in terms of running time, accuracy, and GPU memory consumption comparisons. The comprehensive statistics results are shown in Fig. 5. As can be seen, the overall reconstruction quality of our method is better than MVSNet and R-MVSNet. In addition, the GPU memory consumption of our method is decreased and the running time is also shortened. Compared with Fast-MVSNet, our method uses 12% less memory and 42% less run-time. This is mainly due to the two-stage cost volume construction being constructed at low resolution, while the performance is comparable to these state-of-the-art methods.

3) Evaluation on Outdoor Tanks and Temples Dataset: In the following, to evaluate the generalization ability of the proposed method on the Tanks and Temples datasets without any fine-tuning, the F-score is selected to measure the overall quality of the reconstructed point cloud. In addition, the number of source views is set to 5. Quantitative comparison results with other methods are illustrated in Table II, which indicates that the reconstructed quality of our method is better than Point-MVSNet [13] and Fast-MVSNet in all eight testing scenes. In addition, the average F-score of our method is higher than other methods. Fig. 6 shows the qualitative result of our method, which can achieve dense and continuous 3D scene reconstruction, and the details are smoother and finer even in low-textured regions. The experimental results also indicate that although our network architecture is simple, our model demonstrates comparable performance with limited memory bottleneck and real-time requirements.

D. Ablation Study

1) Quantitative Ablation Analysis: Table III shows the detailed quantitative comparisons of the proposed cost aggregation module. The average absolute depth error metric and prediction accuracy metric within a fixed distance threshold (2mm and 4mm are listed) are selected to measure the quality

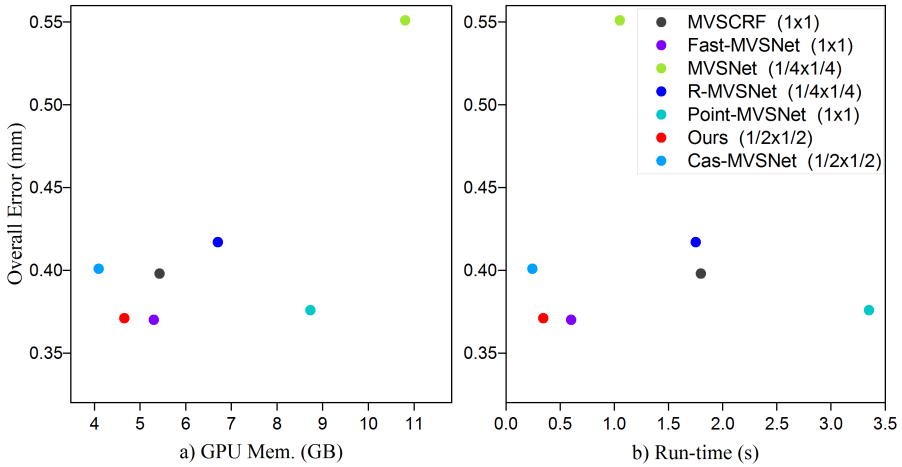


Fig. 5. Comparison results of a) GPU memory consumption and b) running time on DTU datasets.

TABLE II

QUANTITATIVE COMPARISONS OF SCENE RECONSTRUCTION QUALITY OF TANKS AND TEMPLES DATASETS

Methods	Mean	Train	Playground	Horse	Lighthouse	M60	Panther	Francis	Family
COLMAP [23]	42.14	42.04	48.53	25.63	56.43	44.83	46.97	22.25	50.41
Point-MVSNet [13]	48.27	43.06	52.38	34.20	50.79	51.97	50.85	41.15	61.79
MVSNet [9]	43.48	34.69	47.90	25.07	50.09	53.96	50.86	28.55	55.99
MVSCRF [39]	45.73	39.68	52.60	29.93	51.15	50.61	51.45	30.60	59.83
R-MVSNet [10]	48.40	42.38	52.00	32.59	42.95	51.88	48.80	46.65	69.96
Fast-MVSNet [11]	47.39	42.91	53.27	34.98	47.81	49.16	46.2	39.59	65.18
Ours	**53.10**	**50.56**	**55.04**	**35.70**	**55.16**	**60.48**	**53.34**	45.63	68.89

Fig. 6. Point cloud reconstruction results of our model on Tanks and Temples datasets.

TABLE III

COMPARISON RESULTS OF DIFFERENT COST AGGREGATION METHODS ON DTU DATASETS

Methods	Resolution	Mean abs error	<2mm	<4mm	<8mm
MVSNet [9]	$1/4 \times 1/4$	11.63	63.10%	79.95%	87.24%
MVSNet [9] +AM	$1/4 \times 1/4$	9.46	67.45%	80.78%	88.20%
Our Baseline	$1/2 \times 1/2$	7.12	77.19%	86.62%	91.38%
Ours + AM	$1/2 \times 1/2$	**5.94**	**81.09%**	**89.18%**	**92.75%**

of the depth map. It can be seen that after adding the proposed aggregation mechanism, the performance of MVSNet and our baseline have been improved, which shows that the proposed cost aggregation mechanism can aggregate reliable costs and

TABLE IV
PERFORMANCE EVALUATION OF SURFACE NORMAL ON DTU DATASETS

Methods	Error			Accuracy		
	mean	median	rmse	11.25°	22.5°	30°
FCN8 [17] (VGG based)	36.31	31.56	35.86	10.60	40.15	56.33
SEGNET [40] (VGG based)	40.59	35.83	40.20	9.53	29.47	45.11
Ours	23.21	18.37	22.96	15.57	55.01	67.17
Ours+refine	22.27	17.16	21.98	19.03	60.33	71.41
Ours+Edge-refine	**22.18**	**17.04**	**21.86**	**19.12**	**60.52**	**71.49**

[1] VGG stands for VGG-16 backbone.
[2] FCN8 and SEGNET also use VGG as the basic backbone.

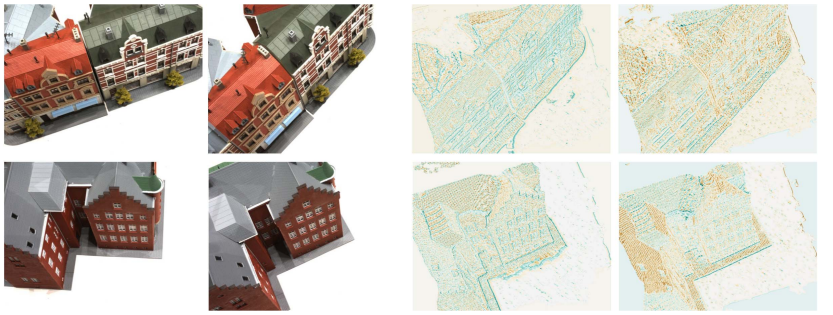


Fig. 7. Qualitative comparisons of the cost aggregation module between source and reference image.

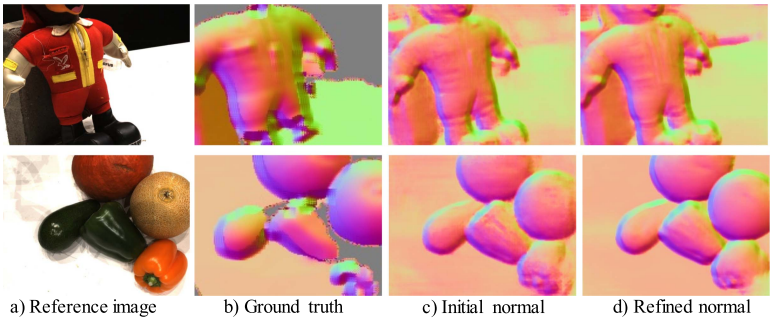


Fig. 8. Qualitative ablation studies for surface normal.

effectively suppress noise. Table IV illustrates the performance comparisons of surface normal estimation. Compared with other networks that predict surface normal through a single image, it can be observed that the cost regularization along the sampling depth direction can predict surface normal with higher accuracy under the same epochs. It can be clearly seen that the accuracy of normal estimation can be further enhanced by the proposed refinement module, which is conducive to reliable cost aggregation. In addition, by adding an edge image in the refinement module, the smoothness consistency constraint can slightly improve the quality of surface normal.

2) Qualitative Ablation Analysis: Fig. 7 shows the qualitative analysis of the adaptive cost aggregation module. Note that the co-visible areas indicate that they can be seen in the source and reference views. It can be seen that after adding the cost aggregation module, the pixels with continuous depth and their neighboring pixels become more visible and smooth. In addition, it can be concluded that the adaptive cost aggregation can describe the potential pixel-visibility relationship even in the non-textured regions, and can produce more accurate co-visible areas. Qualitative results of the surface normal are shown in Fig. 8. Benefiting from the learnable edge-refine module, the predicted surface normal appears more accurate and maintains finer details. It is worth noting that the predicted surface normal map is even more complete than the ground truth. Although the neighboring pixel-visibility can be characterized without using the high-precision surface normal, the accurate normal estimation based on the edge-refine module is more conducive to the adaptive cost aggregation, and the predicted depth is more robust and effective. Fig. 9 shows the qualitative comparison of depth maps. Note that our two-stage network architecture is similar to that of the CasMVSNet. For a fair comparison, we only compare our proposed model with the two-stage depth inference model provided by CasMVSNet. The output resolution of its depth map is also half of the input. It can be seen that our cost aggregation module can make the boundary prediction more accurate and maintain finer details

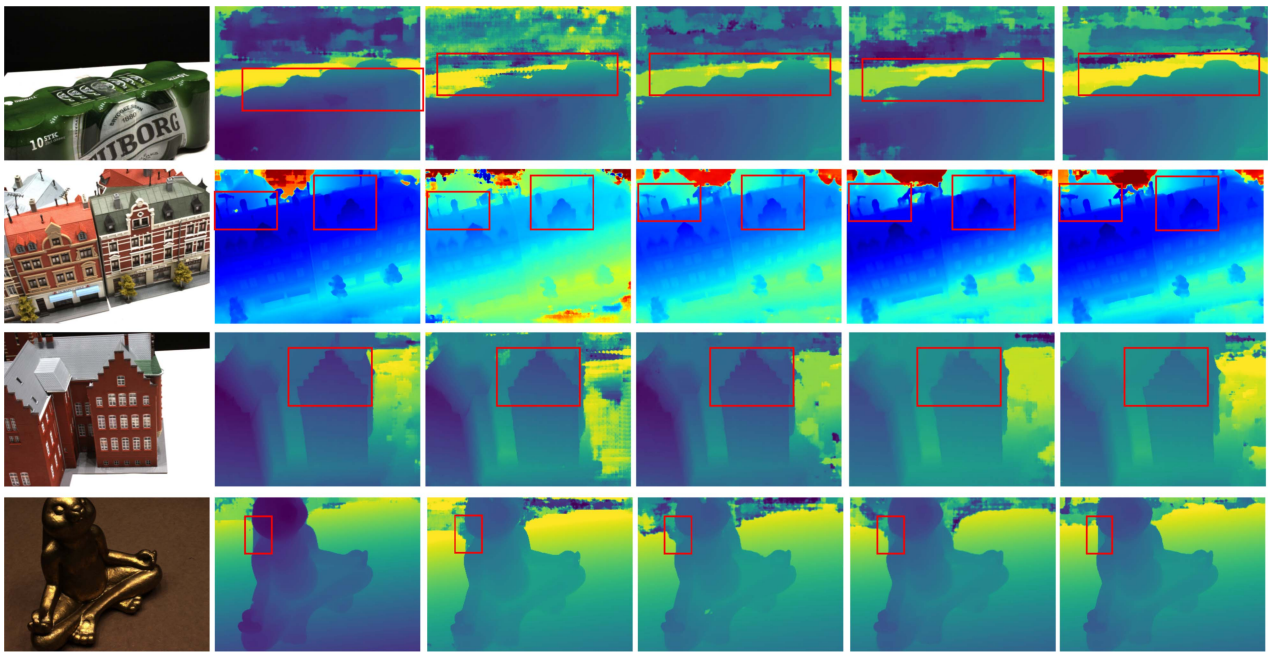


Fig. 9. Qualitative comparisons of depth maps generated by different methods on DTU datasets.

(the red rectangle areas in Fig. 9(f). The quality of depth estimation using the aggregation module is higher than for other methods.

V. Conclusion

In this work, a two-stage multi-view stereo network for fast and accurate depth inference is proposed. First, sparse costs are constructed to jointly estimate coarse depth maps and surface normal with low memory consumption. An edge refinement block is designed to improve the accuracy of surface normal prediction. Second, since the surface normal represents more local geometric property, the similarity of the neighboring surface normal is directly leveraged as the weighted attention strategy to characterize the pixel-visibility, rather than designing an independent module to learn the pixel-visibility for cost aggregation. This strategy makes the depth estimation more accurate and validates that the visual visibility cues contained in the surface normal is more robust and anti-noise. Although the network architecture of the proposed model is relatively simple, comprehensive experimental comparisons on the DTU and the Tanks and Temples datasets show that our method can achieve low memory consumption, favorable generalization properties, and competitive performance compared with other methods. Our method is the first depth estimation framework that leverages surface normal to capture the pixel-visibility for cost aggregation and is compatible with other depth/normal prediction frameworks to achieve 3D reconstruction. In future work, we would like to improve the precision of surface normal estimation based on a multi-view network, remove the restriction of the pre-calibrated camera parameters, and apply the proposed model to unsupervised applications.

References

[1] D.-D. Nguyen, A. Elouardi, S. A. R. Florez, and S. Bouaziz, "HOOFR SLAM system: An embedded vision SLAM algorithm and its hardware-software mapping-based intelligent vehicles applications," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 11, pp. 4103–4118, Nov. 2019.

[2] J. Leng, Y. Liu, D. Du, T. Zhang, and P. Quan, "Robust obstacle detection and recognition for driver assistance systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 4, pp. 1560–1571, Apr. 2020.

[3] W. Chuah, R. Tennakoon, R. Hoseinnezhad, D. Suter, and A. Bab-Hadiashar, "Semantic guided long range stereo depth estimation for safer autonomous vehicle applications," *IEEE Trans. Intell. Transp. Syst.*, early access, May 9, 2022, doi: 10.1109/TITS.2022.3170870.

[4] S. Duggal, S. Wang, W.-C. Ma, R. Hu, and R. Urtasun, "DeepPruner: Learning efficient stereo matching via differentiable PatchMatch," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4384–4393.

[5] J. Sun, N.-N. Zheng, and H.-Y. Shum, "Stereo matching using belief propagation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 7, pp. 787–800, Jul. 2003.

[6] H. Aanæs, R. R. Jensen, G. Vogiatzis, E. Tola, and A. B. Dahl, "Large-scale data for multiple-view stereopsis," *Int. J. Comput. Vis.*, vol. 120, no. 2, pp. 153–168, 2016.

[7] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun, "Tanks and temples: Benchmarking large-scale scene reconstruction," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–13, Jul. 2017.

[8] B. Lu *et al.*, "Toward image-guided automated suture grasping under complex environments: A learning-enabled and optimization-based holistic framework," *IEEE Trans. Autom. Sci. Eng.*, early access, Dec. 29, 2021, doi: 10.1109/TASE.2021.3136185.

[9] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "MVSNet: Depth inference for unstructured multi-view stereo," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 767–783.

[10] Y. Yao, Z. Luo, S. Li, T. Shen, T. Fang, and L. Quan, "Recurrent MVSNet for high-resolution multi-view stereo depth inference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5520–5529.

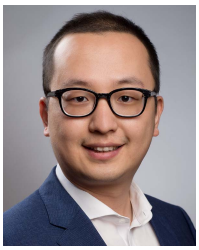
[11] Z. Yu and S. Gao, "Fast-MVSNet: Sparse-to-dense multi-view stereo with learned propagation and gauss-Newton refinement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1946–1955.

[12] S. H. Im, H. G. Jeon, and S. Lin, "DPSNet: End-to-end deep plane sweep stereo," 2019, *arXiv:1905.00538*.

[13] R. Chen, S. F. Han, J. Xu, and H. Su, "Point based multi-view stereo network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Jul. 2019, pp. 1538–1547.
[14] F. Wang, S. Galliani, C. Vogel, P. Speciale, and M. Pollefeys, "PatchmatchNet: Learned multi-view patchmatch stereo," 2020, *arXiv:2012.01411*.
[15] H.-C. Yang, P.-H. Chen, K.-W. Chen, C.-Y. Lee, and Y.-S. Chen, "FADE: Feature aggregation for depth estimation with multi-view stereo," *IEEE Trans. Image Process.*, vol. 29, pp. 6590–6600, 2020.
[16] H. Yi *et al.*, "Pyramid multi-view stereo net with self-adaptive view aggregation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 766–782.
[17] X. Qi, Z. Liu, R. Liao, P. H. S. Torr, R. Urtasun, and J. Jia, "GeoNet++: Iterative geometric neural network with edge-aware refinement for joint depth and surface normal estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 2, pp. 969–984, Feb. 2022.
[18] A. Saxena, S. H. Chung, and A. Y. Ng, "Learning depth from single monocular images," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 1161–1168.
[19] Y. Li and Z. Wang, "RGB line pattern-based stereo vision matching for single-shot 3-D measurement," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.
[20] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2650–2658.
[21] X. Wang, D. F. Fouhey, and A. Gupta, "Designing deep networks for surface normal estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 539–547.
[22] S. Galliani, K. Lasinger, and K. Schindler, "Massively parallel multiview stereopsis by surface normal diffusion," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 873–881.
[23] J. L. Schonberger and J.-M. Frahm, "Structure-from-Motion revisited," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4104–4113.
[24] Q. Xu and W. Tao, "Multi-scale geometric consistency guided multi-view stereo," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5478–5487.
[25] C. Yang, M. Jiang, M. Chen, M. Fu, J. Li, and Q. Huang, "Automatic 3-D imaging and measurement of human spines with a robotic ultrasound system," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.
[26] M. Q. Ji, J. Gall, H. T. Zheng, Y. B. Liu, and L. Fang, "SurfaceNet: An end-to-end 3D neural network for multiview stereopsis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2307–2315.
[27] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 8, pp. 1362–1376, Aug. 2009.
[28] Y. Zhang and T. Chen, "Efficient inference for fully-connected CRFs with stationarity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 109–117.
[29] X. Gu, Z. Fan, S. Zhu, Z. Dai, F. Tan, and P. Tan, "Cascade cost volume for high-resolution multi-view stereo and stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2492–2501.
[30] S. Cheng *et al.*, "Deep stereo using adaptive thin volume representation with uncertainty awareness," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2521–2531.
[31] Q. Xu and W. Tao, "PVSNet: Pixelwise visibility-aware multi-view stereo network," 2020, *arXiv:2007.07714*.
[32] E. Tola, C. Strecha, and P. Fua, "Efficient large-scale multi-view stereo for ultra high-resolution image sets," *Mach. Vis. Appl.*, vol. 23, no. 5, pp. 903–920, May 2011.
[33] J. Zeng *et al.*, "Deep surface normal estimation with hierarchical RGB-D fusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6153–6162.
[34] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of RGB-D SLAM systems," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot Syst. (IROS)*, Oct. 2012, pp. 573–580.
[35] H. Xu and J. Zhang, "AANet: Adaptive aggregation network for efficient stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1956–1965.
[36] R. Jensen, A. Dahl, G. Vogiatzis, E. Tola, and H. Aanaes, "Large scale multi-view stereopsis evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 406–413.
[37] P. H. Huang, K. Matzen, J. Kopf, N. Ahuja, and J. B. Huang, "DeepMVS: Learning multi-view stereopsis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2821–2830.
[38] S. Galliani, K. Lasinger, and K. Schindler, "Massively parallel multiview stereopsis by surface normal diffusion," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 873–881.
[39] Y. Xue *et al.*, "MVSCRF: Learning multi-view stereo with conditional random fields," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4312–4321.
[40] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.
[41] A. O. Ulusoy, M. J. Black, and A. Geiger, "Semantic multi-view stereo: Jointly estimating objects and voxels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 539–547.
[42] Q. Sun, J. Yuan, X. Zhang, and F. Duan, "Plane-edge-SLAM: Seamless fusion of planes and edges for SLAM in indoor environments," *IEEE Trans. Autom. Sci. Eng.*, vol. 18, no. 4, pp. 2061–2075, Oct. 2021.
[43] H. Gao, J. Xiao, Y. Yin, T. Liu, and J. Shi, "A mutually supervised graph attention network for few-shot segmentation: The perspective of fully utilizing limited samples," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Mar. 14, 2022, doi: 10.1109/TNNLS.2022.3155486.
[44] H. Gao, K. Xu, M. Cao, J. Xiao, Q. Xu, and Y. Yin, "The deep features and attention mechanism-based method to dish healthcare under social IoT systems: An empirical study with a hand-deep local-global net," *IEEE Trans. Computat. Social Syst.*, vol. 9, no. 1, pp. 336–347, Feb. 2022.
[45] J. Xiao, H. Xu, H. Gao, M. Bian, and Y. Li, "A weakly supervised semantic segmentation network by aggregating seed cues: The multi-object proposal generation perspective," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 17, no. 1s, pp. 1–19, Mar. 2021, doi: 10.1145/3419842.
[46] J. Chen *et al.*, "A transfer learning based super-resolution microscopy for biopsy slice images: The joint methods perspective," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 1, pp. 103–113, Feb. 2021.
[47] T. Chen *et al.*, "Discriminative cervical lesion detection in colposcopic images with global class activation and local bin excitation," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 4, pp. 1411–1421, Apr. 2022.
[48] Q. Cai, L. Zhang, Y. Wu, W. Yu, and D. Hu, "A pose-only solution to visual reconstruction and navigation," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Dec. 31, 2021, doi: 10.1109/TPAMI.2021.3139681.
[49] G. Zhu, X. Huang, R. Yang, and R. Sun, "Relationship extraction method for urban rail transit operation emergencies records," *IEEE Trans. Intell. Vehicles*, early access, Mar. 25, 2022, doi: 10.1109/TIV.2022.3160502.
[50] E. Q. Wu *et al.*, "ROpenPose: A rapider OpenPose model for astronaut operation attitude detection," *IEEE Trans. Ind. Electron.*, vol. 69, no. 1, pp. 1043–1052, 2022.
[51] E. Q. Wu *et al.*, "Scalable gamma-driven multilayer network for brain workload detection through functional near-infrared spectroscopy," *IEEE Trans. Cybern.*, early access, 2021.
[52] E. Q. Wu *et al.*, "Brain through brain cognitive map and multi-layer latent incremental learning model," *IEEE Trans. Cybern.*, early access, 2021.

Wei Tong received the M.S. degree from the School of Mechanical Engineering, Jiangsu University of Science and Technology, Zhenjiang, China, in 2017. He is currently pursuing the Ph.D. degree with the School of Mechanical Engineering, Nanjing University of Science and Technology. His research interests include SLAM, 3D reconstruction, and machine learning.

Xiaorong Guan received the Ph.D. degree in mechanical engineering from the Nanjing University of Science and Technology, Nanjing, China, in 2008. From 2009 to 2011, he was a Post-Doctoral Researcher with Tsinghua University. From 2018 to 2019, he was a Visiting Scholar with The University of Electro-Communications. He is currently an Associate Professor with the School of Mechanical Engineering, Nanjing University of Science and Technology. He is also a Standing Member of the Wearable Technology Professional Committee, Chinese Institute of Command and Control. He has published over 50 research papers as the first and the corresponding author in international journals and conferences. His research interests include exoskeleton robots, special equipment for mechatronics, and computer vision.



Jian Kang (Member, IEEE) received the B.S. and M.E. degrees in electronic engineering from the Harbin Institute of Technology (HIT), Harbin, China, in 2013 and 2015, respectively, and the Dr.-Ing. degree from the Signal Processing in Earth Observation (SiPEO), Technical University of Munich (TUM), Munich, Germany, in 2019. In August 2018, he was a Guest Researcher at the Institute of Computer Graphics and Vision (ICG), TU Graz, Graz, Austria. From 2019 to 2020, he was with the Faculty of Electrical Engineering and Computer Science, Technische Universität Berlin (TU Berlin), Berlin, Germany. He is currently with the School of Electronic and Information Engineering, Soochow University, Suzhou, China. His research focuses on signal processing and machine learning techniques and their applications in remote sensing. He has obtained the First Place of the Best Student Paper Award in EUSAR 2018, Aachen, Germany. His joint work was selected as one of the ten Student Paper Competition Finalists in IGARSS 2020.

Poly Z. H. Sun (Member, IEEE) is currently pursuing the Ph.D. degree in industrial intelligent system with the Department of Industrial Engineering, School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai, China.

He is a Research Assistant with the Department of Automation, Shanghai Jiao Tong University. He has authored/coauthored over 20 research papers in top-tier refereed international journals and conferences. His current research interests include intelligent transportation systems, evolutionary optimization, neuroergonomics, brain and cognitive science, complex networks, non-parametric machine learning, operations research, and their applications in industrial or medical problems. He is a member of the IEEE CIS, IEEE SMC, and Association for Computing Machinery. He has been severing as a reviewer or the session chair for several top-tier international journals and conferences in his research field.

Rob Law is the University of Macau Development Foundation (UMDF) Chair Professor of smart tourism. He is also an Honorary Professor of several other reputable universities. Prior to joining the University of Macau in July 2021, he has worked in industry organizations and academic institutes. He is an Active Researcher. He has received more than 90 research related awards and accolades, as well as millions of USD external and internal research grants. He has edited four books and published more than 1,000 research papers (including hundreds of articles in first-tier academic journals). His publications have received more than 53,500 citations, with H-index/i 10-index = 105/478 (http://scholar.google.com.hk/). In addition, he serves different roles for more than 200 research journals. He is the chair/committee member of more than 180 international conferences.

Pedram Ghamisi (Senior Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of Iceland, Reykjavik, Iceland, in 2015.

Since 2018, he has been working as the Head of the Machine Learning Group, Helmholtz-Zentrum Dresden-Rossendorf (HZDR). He is also the CTO and the Co-Founder of VasoGnosis Inc., Milwaukee, WI, USA, where he is involved in the development of advanced diagnostic and analysis tools for brain diseases using cloud computing and deep learning algorithms. His research interests include interdisciplinary research on remote sensing and machine (deep) learning, image and signal processing, and multisensor data fusion. He was a recipient of the Best Researcher Award for M.Sc. students at the K. N. Toosi University of Technology in the academic year of 2010–2011, the IEEE Mikio Takagi Prize for winning the Student Paper Competition at IEEE International Geoscience and Remote Sensing Symposium (IGARSS) in 2013, the Talented International Researcher by Iran's National Elites Foundation in 2016, the First Prize of the Data Fusion Contest organized by the Image Analysis and Data Fusion Technical Committee (IADF) of IEEE-GRSS in 2017, the Best Reviewer Prize of IEEE GEOSCIENCE AND REMOTE SENSING LETTERS in 2017, the IEEE Geoscience and Remote Sensing Society 2019 Highest Impact Paper Award, the Alexander von Humboldt Fellowship from the Technical University of Munich, and the High Potential Program Award from HZDR. He is the Vice-Chair of the IEEE Image Analysis and Data Fusion Committee. He serves as an Associate Editor for *Remote Sensing* (MDPI), *Sensors* (MDPI), and IEEE GEOSCIENCE AND REMOTE SENSING LETTERS.

Edmond Q. Wu (Senior Member, IEEE) received the Ph.D. degree in controlling theory and engineering from Southeast University, Nanjing, China, in 2009.

He is currently a Professor with the Key Laboratory of System Control and Information Processing, Ministry of Education, Shanghai Jiao Tong University, China. His research interests include deep learning, fatigue recognition, and human–machine interaction. He is an Associate Editor of IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS and IEEE TRANSACTIONS ON INTELLIGENT VEHICLES. He is also a Guest Editor of IEEE TRANSACTIONS ON COGNITIVE AND DEVELOPMENTAL SYSTEMS.