# Multistage Pixel-Visibility Learning With Cost Regularization for Multiview Stereo

Xiaorong Guan, Wei Tong, Shan Jiang, Poly Z. H. Sun, *Member, IEEE*, Edmond Q. Wu, *Senior Member, IEEE*, and Guimin Chen

***Abstract*— Multiple-view stereo has potential applications in robotic operations and autonomous driving (unstructured environment construction, visual servo). With assisted depth information, inertial navigation systems can achieve precise navigation. It is, especially suitable for GPS failures in complex environments. Accurate depth estimation is a challenge in low-textured or occluded regions. To alleviate the inference of incorrect depth, a multi-stage pixel-visibility learning-based stereo network is presented in this paper. Its improvements are as follows: 1) a new content-adaptive cost volume aggregation mechanism based on neighboring pixel-wise visibility is designed to effectively produce more accurate and smoother depth map predictions in the object boundary. 2) global convolution block and boundary refinement block are developed to regularize its cost volume, they can learn the inherent constraints of feature matching correspondence and effectively mitigate the depth estimation uncertainty in low-textured regions. 3) a new loss function is designed to measure the uncertainty of predicted probability distribution and enhance the reliability of depth map inference. Experimental results on the indoor DTU datasets and the outdoor Tanks & Temples datasets indicate that our method can achieve superior performance and has a powerful generalization ability, which is comparable to state-of-the-art works.**

Manuscript received January 8, 2022; revised March 4, 2022; accepted April 2, 2022. This article was recommended for publication by Associate Editor Q. Xu and Editor L. Zhang upon evaluation of the reviewers' comments. This work was supported in part by the National Defense Basic Scientific Research Program of China under Grant JCKY2019209B003; in part by the National Natural Science Foundation of China through the Main Research Project on Machine Behavior and Human–Machine Collaborated Decision Making Methodology under Grant 72192820; in part by the Third Research Project on Human Behavior in Human–Machine Collaboration under Grant 72192822; and in part by the National Natural Science Foundation of China under Grant U1913213, Grant 62171274, and Grant U1933125. *(Corresponding authors: Wei Tong; Shan Jiang; Poly Z. H. Sun.)*

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

Shan Jiang is with Johnson & Johnson Supply Chain, Bridgewater Township, NJ 08933 USA (e-mail: sj576@scarletmail.rutgers.edu).

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

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@sjtu.edu.cn).

Guimin Chen is with the State Key Laboratory of Manufacturing Systems Engineering and the Shaanxi Key Laboratory of Intelligent Robots, School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: guimin.chen@xjtu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TASE.2022.3165944.

Digital Object Identifier 10.1109/TASE.2022.3165944

***Note to Practitioners*— Multiple-view stereo (MVS) can estimate dense 3D representations of scenes, which is widely used in autonomous driving, robotic navigation, virtual reality (VR), and augmented reality (AR). Aiming at the problem of incorrect depth inference in low-textured or occluded regions, this work proposes a novel multi-stage depth prediction method based on neighboring pixel-wise visibility. Our method cannot only achieve accurate depth estimation for robot perception but also make no concession to real-time performance. It is clear that the proposed method has good potential in 3D reconstruction, robotic navigation, and VR/AR fields to provide accurate depth estimation in real-time with limited memory consumption.**

***Index Terms*— Cost aggregation, depth prediction, multiview stereo, global convolution, pixel-aware visibility.**

## I. Introduction

With the broad prospects in various tasks such as 3D reconstruction, machine-assisted surgery and autonomous driving, multiple-view stereo (MVS) has attracted widespread attention in both academia and industry. As a fundamental computer vision problem for several years, it aims to estimate the depth maps and further reconstruct dense 3D point clouds through multiple image sequences of scenes with overlapping regions, including important values for the applications of virtual reality, robotics, and autonomous driving. Since depth-based methods are not susceptible to geomagnetic interference and do not require additional electronic devices, they are widely used in the positioning system and navigation tasks. For example, Zhang *et al.* [1] introduced the 3D vision information to determine the position and orientation of the drogue. Hussain *et al.* [2] proposed a stereo visual-inertial tracking for lower limb tracking in physiotherapy applications. Other navigation methods [3], [39] have been proposed to perceive the environment and localize the robot by assisted monocular and binocular vision. Traditional MVS methods [4]–[7] are mainly based on hand-crafted features for similarity matching and can achieve the average performance of 3D scene reconstruction. However, these types of methods cannot achieve reliable and accurate feature matching correspondences under the conditions of illumination changes, specular reflection, and low-textured regions. Benefiting from the success of convolutional neural networks (CNN) and continuous improvement of 3D reconstruction datasets, recent learning-based methods [8]–[10] can achieve good performance. The deep learning network extracts more discriminative image features and implicitly encodes the local prior geometry such as specular

reflection to construct the 3D cost volume, and further regularizes to estimate the final depth. As for learning-based methods, the feature matching correspondences from multiple images are more robust, which can remarkably improve the overall performance. Yu *et al.* [16] introduce the attention-aware cost pyramid module to improve the quality of 3D reconstruction. Giang *et al.* [17] enhance the performance of feature matching by learning a robust feature extraction network without heavy computation. In particular, Yao *et al.* propose MVSNet [15] to estimate the depth value for the reference view by multiple RGB images, where an essential innovation in MVSNet is the construction of corresponding 3D matching volume from the sampled depth hypothesis and the multi-scale 3D CNNS are used to realize the accurate depth inference. Its follow-up work [11]–[14] also adopt the variance-based cost metric to feed multi-view features from all views to suppress the potential noise and interference of the cost volume. However, treating each neighboring image equally will make the cost volume susceptible to noise and cannot guarantee the accuracy of the depth estimation, especially in low-textured regions. In addition, wrong visibility for the cost volume aggregation mechanism may inevitably deteriorate the final reconstruction. Although existing learning-based MVS methods [30], [36] additionally apply the max pooling and averaging operation to generate the pixel-visibility map for multi-scale cost volume aggregation, they are restricted by limited memory and depth precision under noise conditions still need to be further improved.

In this paper, to guarantee reliability and accuracy even in low-textured regions, the trainable multi-view stereo network architecture is proposed as shown in Fig. 1. Firstly, multi-scale image features are extracted based on the feature pyramid. Then cost volume is separately constructed under the standard sampled depth hypotheses in a coarse-to-fine manner. The cost volume at early stages is with a low sparse resolution, and at later stages is to adaptively zoom out and adjust the sampled depth interval with high dense resolution. Secondly, assuming that the depth value of neighboring pixels is highly continuous and correlated, a pixel-wise visibility-aware module is designed to adaptively learn the position-specific weighting for more robust cost volume aggregation and suppress the pixel mismatch in low-textured regions, the cost volume is then regularized to obtain final depth estimation. Thirdly, instead of applying the general 3D CNNs to regularize the cost volume, global 3D CNNs block and boundary refinement block are proposed to further learn the inherent constraints of feature matching correspondence. In addition, to improve the reliability of the depth estimation, the uncertainty loss function is incorporated to regularize the matching probability distribution, which can effectively increase the quality of the overall probability map and enable introduce a more accurate depth map.

Our main contributions are listed as follows: 1) A multi-stage pixel-visibility learning-based multi-view stereo network for accurate depth inference is proposed. An encoder-decoder module is designed to learn each two-view pixel-visibility map, which can reflect the influence of occlusion, illumination, and unstructured viewing geometry. 2) A self-adaptive pixel-wise visibility-aware cost volume aggregation mechanism is proposed, which can aggregate reliable cost volume even in low-textured regions and object boundaries. 3) A new cost volume regularization module is applied to process the aggregated cost volume, which focuses on suppressing feature mismatch. 4) A novel loss function is designed, which can effectively improve the reliability of the depth inference.

## II. Related Work

### A. Traditional MVS

Traditional MVS methods usually apply the photo-consistency constraints to optimize the depth estimation. Depth-based methods have been widely used to reconstruct point clouds, which can be relatively easy to optimize with flexibility and perform more concisely. The current advanced MVS methods mainly predict the accurate depth value by selecting neighboring pixels, multi-scale cost volume aggregation modules, and applying the local propagation. Previous work that explicitly quantifies the cost volume contributions from multiple views is carried out. Zheng *et al.* [23] estimate visibility and geometry by designing a probabilistic graphical model. Following this framework, Schonberger *et al.* proposed a MVS method named COLMAP [6] that considers the various photometric and geometric priors to better depict their graphical model. COLMAP has become the standard of the traditional MVS methods because of its high accuracy on the benchmarks. However, the entire implementation process is time-consuming and cannot be optimized in parallel. Xu *et al.* [19]. leveraged the multi-hypothesis joint view selection mechanism and applied the adaptive checkerboard propagation to improve the performance of scene reconstruction. These iterative works utilize the difference between neighboring views and consider predefined criteria, while it is hard to be adaptable in other scenarios.

### B. Learning-Based MVS

Traditional MVS methods learn scene geometry by introducing the assumption of photo consistency. However, these types of methods perform poorly in occluded and low-textured regions. The recent increasing amount of studies on learning-based CNNs has tended to exploit MVS without using traditional hand-crafted image features. Learning-based MVS can be divided into the volumetric representation and plane-sweep algorithm [21], plane-sweep methods estimate the depth maps as the intermediate representations. However, compared with plane-sweep methods, the resolution of the reconstructed model based on voxel cubes requires huge memory requirements and easily leads to high time complexity. Moreover, the manner of the volumetric representation method has been empirically proved to be inferior to the depth-based method. To realize the end-to-end training for depth estimation, Yao *et al.* [15] implicitly encoded the 3D geometry into the network by adopting differential homography warping operation, which realizes accurate depth estimation. Furthermore, Yao *et al.* introduced R-MVSNet [9] to replace the original 3D cost space with recursive cost

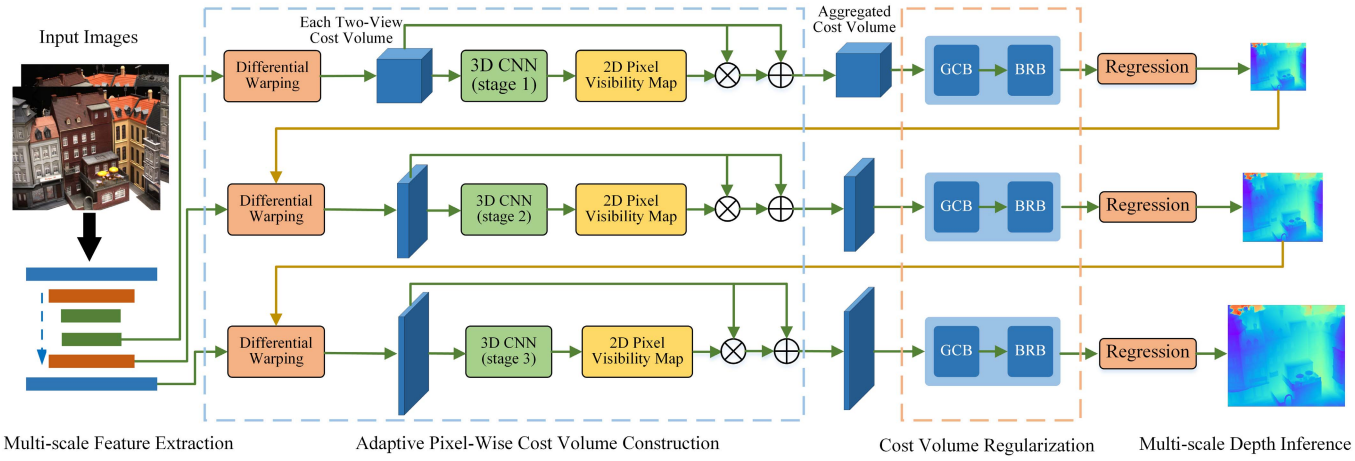



Fig. 1. The overall architecture of our proposed multi-view depth inference network. The architecture consists of multi-scale feature extraction, generation of per-two-view visibility map, adaptive cost volume construction, cost volume regularization, and depth inference.

volume regularization, which is specially designed for possible high-resolution stereo reconstruction with limited memory bottleneck. Point-MVSNet [11] proposed a Point Flow module to refine the depth map through an iterative optimization manner. Thereafter, an increasing number of works [12], [13] related to the weakening of memory and time consumption has been proposed. Gu *et al.* [13] proposed the cascade idea to separately estimate the depth map in a multi-scale manner, which not only guarantees the accuracy but also greatly reduces the memory of cost volume. Cheng *et al.* proposed UCSNet [12] by utilizing variance-based uncertainty estimates to adaptively construct thin volumes in different ranges. Particularly, these depth estimation methods follow the philosophy that each perspective contributes equally to the construction of cost volume, neglecting the fact that the neighboring pixels' depth values are usually continuous and highly correlated even under the low-textured and illuminated regions.

## III. Proposed Method

The architecture of the proposed depth estimation network is now described in detail, which is an effective supplement to the previous work on multi-view stereo methods. The proposed method extracts the image features by the CNN operation and adapts arbitrary views as inputs for the multi-view stereo. The overall network architecture is shown in Fig. 1 and is introduced in the following sections.

### A. Pyramid Feature Extraction

Previous work [18], [22], [24]–[26] mainly extracted the high-level semantic feature maps to construct the standard cost volume by the homography warping while lacking the low-level finer representations. As schematically shown in Fig. 1, to make the multi-view matching among images more robust to the illumination and specular reflection conditions, a small 2D UNet module is applied to realize the encoding and decoding of image semantic information. The size of the feature map at multiple stages is $\{1/4, 1/2, 1\}$ of the input image, respectively. Followed by the method of cost volume construction in MVSNet, three scales of corresponding cost volumes are built on the feature maps. Through the learned up-sampling process, the high-resolution features are reasonably incorporate low-resolution information, thus can leads to reasonable high-frequency feature extraction in the multi-stage depth prediction.

### B. Pixel-Wise Visibility-Aware Cost Volume Construction

*1) Each Two-View Cost Volume Construction:* The feature map of each source image is first warped to the reference image view by the camera coordinate conversion, thereby constructing each two-view cost volume at multi-scales. This process is conducted through front-parallel planes at a multi-sampled depth hypothesis. The coordinate mapping is determined by the homography matrix:

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

where $M_i(d)$ means the homography matrix conversion from the source feature map to the reference feature map under the sampled prior depth value $d$. Moreover, $K_i$, $R_i$, $t_i$ represent the intrinsic parameters, relative rotation matrix, and relative translation matrix of the $i^{th}$ view, respectively. $\eta_1$ refers to the camera principle axis of the reference image. To further form a multi-scale cost volume, the hypothetical depth value is uniformly sampled from the inverse depth space and uses the differentiable homography to warp the source feature map via the reference camera planes. Note that the group-wise correlation [27] is utilized to encode the similarity between the reference feature map and each source feature map, each similarity map is named $C_p$ and its size is $D \times H \times W \times C$, where $D$ denotes the sampled number of the depth hypothesis, $H \times W$ is the resolution of the current feature map, $C$ is the supposed channel number of the similarity map.

*2) Adaptive Pixel-Wise Visibility Cost Volume Aggregation:* After obtaining each two-view cost volume, in the previous work [9], [15], [28], there is a constant function that uses the multi-scale variance-based metric to suppress the noise and generate the final cost volume, which considers that all

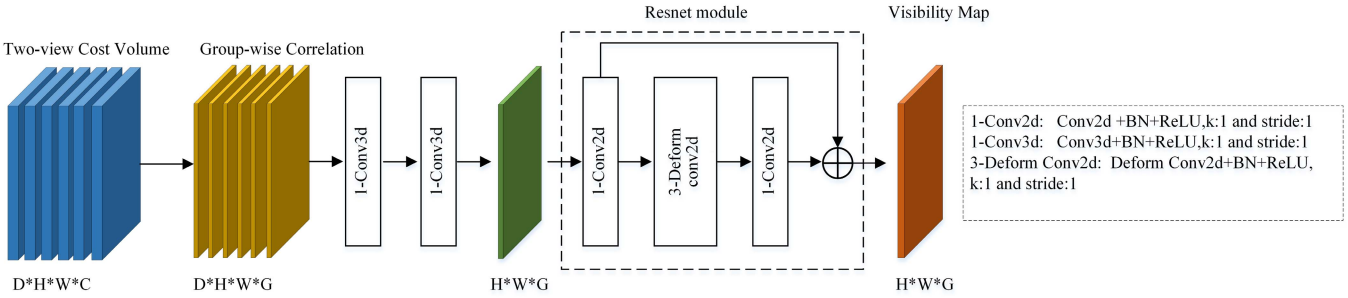



Fig. 2. Pixel-wise visibility-aware weight map module. Each two-view cost volume is first passed through an encoder consisting of group-wise correlation along the depth direction and then decoded by a residual module to obtain the visibility map.

pixels contribute equally to the cost volume. However, this is not reasonable, especially under the occluded areas or illumination regions. Besides, neighboring pixels are usually highly correlated, and the depth values tend to be continuous, while depth values usually vary drastically on the borders of the object.

Inspired by the works proposed in AANet [37] that guide the network to learn the weighted feature matching by the attention mechanism [29], the weighted visibility map is integrated to learn spatial finer representations for cost volume. Different from the works in PatchmatchNet [38] that aggregate spatial cost over the multi-view cost volume, we consider that each pixel-visibility map for per-two-view cost aggregation is similar in the depth dimension of the constructed cost volume, but has different saliency in the dimensions of width and height. Specifically, to decrease the depth estimation errors in low-textured and edge-fattening regions, a self-adaptive pixel-wise visibility-aware weight map module is proposed to learn potential cost volume aggregation, and its detailed architecture is illustrated in Fig. 2. To produce a weighted attention map, 3D CNNs are utilized to encode the matching features of each two-view 3D cost volume in the depth dimension, the size of each two-view processed matching costs are $H \times W \times G$. Then, the cost volume is encoded by a residual connection module and uses deformable convolution instead of standard convolution to enhance the modeling geometric transformation abilities of the self-attention mechanism, the proposed self-adaptive pixel-wise visibility map is defined as:

$$V_i(p) = \frac{\sum_{k=1}^{K^2} \left( \omega_k \cdot m_k \cdot e^{-\alpha_1 |\nabla I_K|} \cdot C_p(p + p_k + \Delta p_k) \right)}{\sum_{k=1}^{K^2} \omega_k \cdot m_k \cdot e^{-\alpha_1 |\nabla I_K|}}, \tag{2}$$

where $V_i(p)$ denotes the visibility map of pixel $p$ for each two-view cost volume $Cp$, $K^2$ is the number of neighboring sampling points for pixel $p$ ($K = 3$ in our paper), $\omega_k$ is the attention weight for the $K^{th}$ point, $p_k$ is the fixed offset to $p$ in window. In addition, the offset $\Delta p_k$ is introduced to learn more detailed neighboring locations for the regular sampling pixel $p + p_k$, which enables adaptive sampling of flexible and efficient neighboring pixels for the current pixel $p$. Hence, high-quality results can still be guaranteed in the object boundary and low-textured areas. Considering the regular convolution weights $\{\omega_k\}_{k=1}^{K^2}$ might make the learning content-agnostic, specific weights $\{m_k\}_{k=1}^{K^2}$ (*i.e.*, modulation in [31]) are used to be learned to achieve content-adaptive neighboring pixels visibility for each pixel location $p$. Different from previous work [37], [38], to cope with the problem of depth value discontinuity in object boundary or some irregular surfaces, the weights $e^{-\alpha_1 |\nabla I_K|}$ is proposed to make the depth estimation more conforming to 3D geometry consistency, which implies that high gradient represents the low pixel visibility weight. Note that a separate convolution operation is applied from the input cost volume to obtain the weight $\Delta p_k$ and $m_k$, Eq. (2) is implemented by deformable convolution. Fig. 3 illustrates several sparse pixels are sampled by our method and stayed within the object boundary, alleviating the ambiguity of depth estimation in low-textured areas.

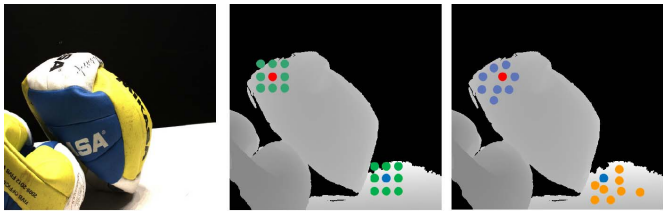
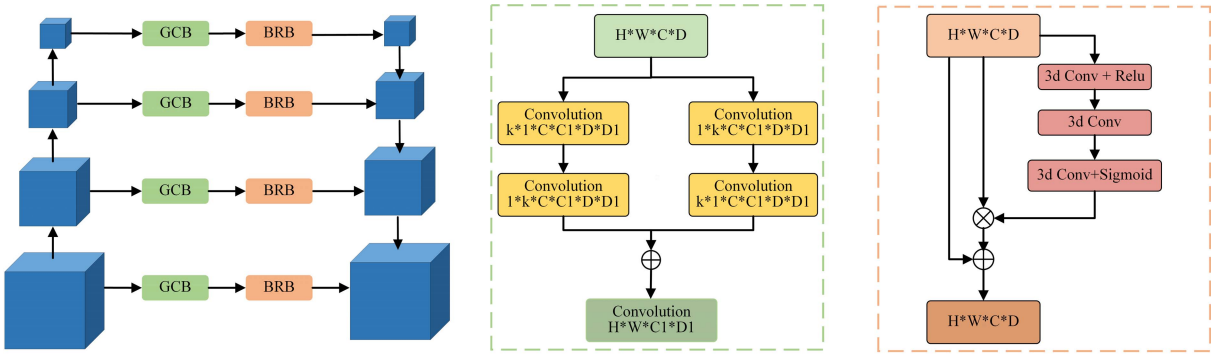


Fig. 3. Comparisons of generally sampled pixels and our adaptive sampled pixels. Left: reference image. Middle: sampled pixels in the fixed window. Right: sampled pixels of the proposed method.

The three convolutional layers with lightweight residual blocks are used to obtain a pixel-wise visibility map, and the convolution kernel sizes are $1 \times 1$, $3 \times 3$ and $1 \times 1$, respectively. Among them, $3 \times 3$ convolution is a deformable convolution. Dilated convolution [32] is also used to make deformable convolution more flexible.

After obtaining the self-adaptive pixel-wise visibility map from each two-view matching cost volume, without leveraging complicated cost metrics in previous work [15], [30] to obtain the final multi-view cost volume, the sum of each two-view cost matching volume is merely averaged, which can already make 3D cost volume regularization more distinguishable. The aggregated cost volume is defined as:

$$C_{agg} = \frac{\sum_{i=1}^{N-1} (1 + V_i(p)) * C_p}{N - 1}. \tag{3}$$




(a) Cost Volume Regularization (b) Global 3D Convolution Block (c) Boundary Refinement Block

Fig. 4. An overview of the proposed cost volume regularization module in (a). The details of the global 3D convolution block and the boundary refinement block are illustrated in (b) and (c).

### C. Cost Volume Regularization and Depth Inference

*1) Cost Volume Regularization:* Influenced by the existence of non-Lambertian surfaces and noise contamination, the aggregated cost volume still needs to be optimized for depth map inference. For cost volume regularization, a 3D UNet module as used in MVSNet is applied to filter the cost volume. Instead of directly using general 3D CNNs to regularize the cost volume, the global 3D convolution block (GCB) is introduced and shown in Fig. 4(b) to extract fruitful semantic features and global geometric attributes. To reduce the computational requirement, our GCB block utilizes the combination of $1 \times k \times k_1 + k \times 1 \times k_1$ and $k \times 1 \times k_1 + 1 \times k \times k_1$ convolutions to replace the general convolution $k \times k \times k_1$. Compared with the general 3D convolution, our GCB block contains only $O(2/k)$ computation cost and can retain a large receptive field in the feature map.

Note that our multi-scale 3D UNet module has four scales, the kernel size decreases as the number of layers deepens, and the kernel size of the four scales are 11, 9, 7, and 5, respectively. To further mitigate the problem of uncertain depth inference in the object boundary, the boundary refinement block (BRB) is also introduced and shown in Fig. 4(c) to add smoothing constraints for depth inference. The BRB consists of a residual structure. More specifically, the $\tilde{P}$ is defined as the refined cost volume: $\tilde{P} = P + P \times \Re(P)$, where $P$ is the cost volume processed by GCB and $\Re(\cdot)$ is the residual branch.

*2) Depth Inference:* After generating the probability volume $P$ through cost volume regularization, the similar depth regression method as described in [15] is used to predict depth value. The final pixel-wise depth value is calculated by the $soft\ argmin$ operation along the depth direction to regress the probability volume, which can be expressed as:

$$D_{est} = \sum_{d=d_{\min}}^{d_{\max}} d \cdot P(d). \tag{4}$$

Note that the $soft\ argmin$ operation is referred to the soft attention mechanism, which can be trained by the backpropagation and is more robust than the $argmax$ operation.

### D. Coarse-to-Fine Network Manner

Our multi-stage depth estimation network architecture is followed by recent work [13], which generally supports different numbers of planes through uniform sampling. In the experiment, the number of sampling depths used to construct the sweep volume is set to 48, 16, and 8, respectively. In the first stage, the predetermined depth range is sampled for the depth values hypothesis with a larger depth interval and the size of the constructed sweep volume is $\frac{W}{4} \times \frac{H}{4} \times 48$. In the following stages, the depth interval becomes smaller and the sampling depth range revolves around the predicted depth value of the previous stage, which already narrows the depth range. The rest size of the constructed cost volume are $\frac{W}{2} \times \frac{H}{2} \times 16$ and $W \times H \times 8$, respectively. Specifically, to check the method of applying the finer hypothesis plane intervals please, refer to [13].

### E. Training Loss Function

In the field of 3D scene reconstruction and monocular depth estimation, most existing works focus on minimizing the discrepancy between ground truth and network output. Different from previous work [8], [15], [33], these works use the $L_1$ loss function to train the network, considering that the probability distribution $\left\{P_{i,j}\right\}_{j=1}^{N_d}$ characterizes the reliability and quality of feature matching to some extent, the standard deviation of the distribution is adopted to measure the uncertainty of predicted probability distribution, and an uncertainty loss is introduced, which is defined as:

$$L_i^{uncertain} = \frac{1}{|I_0^{valid}|} \sum_{x \in I_0^{valid}} e^{\partial_1 \times \delta_i} \left| D_i - D_{gt,i} \right|, \tag{5}$$

where $\delta_i = \sqrt{\sum_{j=1}^{N_d} P_{i,j} \left(D_{i,j} - D_{ave}\right)^2}$ denotes the standard deviation of the probability distribution $\left\{P_{i,j}\right\}_{j=1}^{N_d}$, $D_{ave}$ is the mean value under the depth direction, $I_0^{valid}$ is the valid pixels of the mask, $D_i$ refers to predicted depth value, and $D_{gt,i}$ represents ground truth. On the one hand, after adding the proposed loss function, when the predicted depth value is far from the ground truth, the network is learned to decrease the

standard deviation of probability distribution $\delta_i$ and the abs depth error will penalize the loss function more for model training. On the other hand, when the difference between predicted depth and ground truth gradually shrinks, the model will pay more attention to reducing the probability distribution $\delta_i$. Accordingly, compared with only adding abs depth error, the regularized probability distribution will enhance the overall feature matching quality through the proposed uncertainty loss function, and make it easier for the network to return to a more accurate depth value.

To avoid the proposed uncertainty loss function may over-relax the multi-stage depth value estimation, the $L_1$ loss function is incorporated into the total loss function to ensure the overall depth estimation quality. The final multi-stage loss is expressed as:

$$L = \sum_{K=1}^{3} \lambda_K \left( L_K + L_K^{uncertain} \right) \tag{6}$$

## IV. Experiment Results

In this section, experimental comparison and performance evaluation are conducted with existing state-of-the-art works on the public benchmarks. The evaluation datasets include the indoor DTU datasets [35] and the outdoor Tanks & Temples datasets [34].

### A. Datasets

The large-scale MVS datasets called DTU contain 124 different scenes given camera parameters, every scene has 49 views. The image resolution is 1600 × 1200. The outdoor Tanks & Temples datasets are collected from the real scenes. Compared with DTU datasets, the depth range of each image is smaller. In this paper, its intermediate sets containing 8 scenes are used as the test data to validate the generalization ability of the model.

### B. Implementations

The proposed depth estimation network is trained on DTU datasets. To fairly compare the experiments with other methods, the same training, validation, and testing sets are divided as used in [11], [15], and follow the same data pre-process strategies. In addition, the input images resolution is set to 640 × 512 and the number of source views is set to 3.

For multi-stage depth estimation, the depth hypothesis number is set to 48, 16, and 8, respectively. From the beginning to the final stage, depth map resolution is gradually increased, which is 1/4, 1/2, and 1/1 of the input image resolution. During the training, Adam optimizer is utilized (with $\beta_1 = 0.9$, $\beta_2 = 0.999$) for 20 epochs. The initial learning rate is 0.001 and will be halved at $10^{th}$, $12^{th}$ and $16^{th}$ epochs to avoid falling into the local optima. The proposed network is trained by using one NVidia RTX 3090 GPU and the batch size is 4. For the quantitative evaluation of DTU datasets, the MATLAB code is used to evaluate the completeness and accuracy of the 3D point cloud predicted by the proposed method, the specific evaluation metric is the average value of accuracy and completeness in all testing sets.

TABLE I

Quantitative Comparison Results of Different Methods on DTU Evaluation Sets. Evaluation Manner Is Distance and Unit Is Millimeter. A "Smaller Value" Represents Higher Accuracy

| Methods | Acc. (mm) | Comp. (mm) | Overall |
|---|---|---|---|
| Camp[4] | 0.835 | 0.554 | 0.695 |
| Gipuma[5] | **0.283** | 0.873 | 0.578 |
| SurfaceNet[20] | 0.450 | 1.040 | 0.745 |
| MVSNet[15] | 0.456 | 0.646 | 0.551 |
| MVSCRF [8] | 0.371 | 0.426 | 0.398 |
| Fast-MVSNet[40] | 0.336 | 0.403 | 0.370 |
| R-MVSNet[9] | 0.383 | 0.452 | 0.417 |
| Cas-MVSNet[13] | 0.325 | 0.385 | 0.355 |
| PatchmatchNet[38] | 0.427 | **0.277** | 0.352 |
| Our 1st stage | 0.630 | 0.567 | 0.598 |
| Our 2nd stage | 0.396 | 0.402 | 0.399 |
| Our final model | 0.321 | 0.343 | **0.332** |

### C. Evaluation on DTU Datasets

Comprehensive comparisons are conducted with learning-based methods and traditional methods on the DTU testing set containing 22 scenarios. For a fair comparison, the basic parameter settings in our paper are mostly consistent with the existing works [8], [9], [15], and the output depth map resolution is 1600 × 1184. The quantitative results of each method on the DTU evaluation set are shown in Table I. Note that with the same image resolution, the Point-MVSNet output the depth map with the original image resolution, depth map resolution predicted by MVSNet and R-MVSNet are both one-quarter of the input image, our predicted depth map resolution is as same as the input image. As can be seen in Table I, Gipuma [5] performs well in terms of point cloud accuracy metric compared with other methods, while the overall performance of the point cloud predicted by our proposed method is the best. This might be attributed to the integration of the adaptive pixel-wise visibility mechanism and new cost volume regularization module, which can achieve more precise depth inference to reconstruct a high-quality 3D point cloud. Compared with PatchmatchNet [38], the accuracy metric of our model is remarkably better, which might be due to the strategy of adaptive visibility-aware cost aggregation. In particular, our model in the second stage is already comparable to all methods and our final results outperform all other baseline methods in terms of accuracy and completeness.

The evaluation results of running time, accuracy, and GPU memory consumption are shown in Fig. 5, compared with MVSNet, the total GPU memory consumed by our method is decreased from 10823MB to 6274MB and the run-time of our method is reduced by 2 times. Compared with other methods, the GPU memory consumption is slightly higher than Fast-MVSNet [40], while the overall error of our reconstructed point cloud is preferable to all other methods. Besides, the qualitative comparison results are illustrated in Fig. 6. It can be seen that compared with MVSNet and R-MVSNet, benefiting from the proposed multi-stage stereo network, our method can regress a smoother depth map and the reconstructed point cloud contains finer details. For example, our method generates

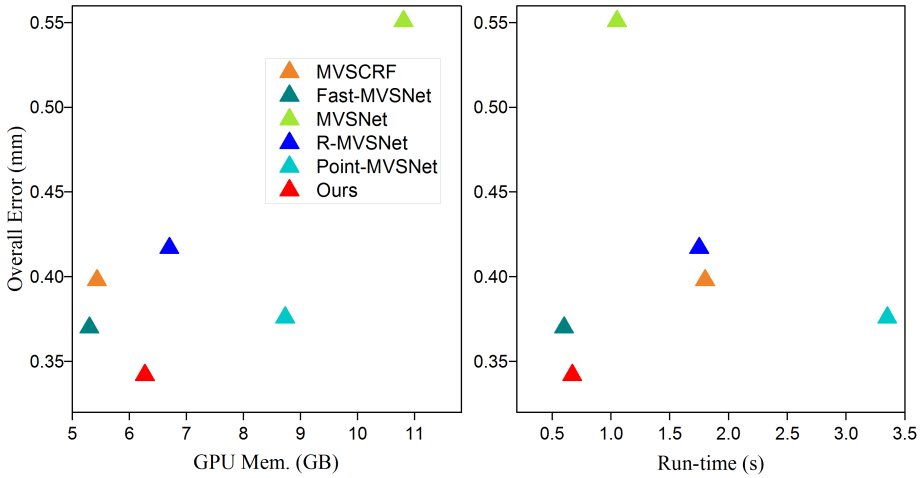



Fig. 5. The comparison results of GPU memory consumption and running time between our model and the methods in Ref. [8], [9], [11], [15], and [40] on the DTU testing set. The size of our predicted depth map is the same as the input image.

Fig. 6. Qualitative comparisons of several scenes point cloud on DTU datasets. Left: MVSNet. Middle: R-MVSNet. Right: Our method. From top to down are scan1, scan62, scan77, scan114 on the testing set. The point clouds clearly show our method can achieve dense and accurate reconstruction, which is even of higher quality than ground truth (ours: vs. MVSNet: 192 and R-MVSNet: >512).

the most complete points clouds even if the object shape is naturally not obvious or low-textured (*e.g.*, top cover and metal base in the third row of Fig. 6). Note that sampled depth hypothesis number of R-MVSNet is set to 512 and 192 in MVSNet.

### *D. Evaluation on Outdoor Tanks & Temples Dataset*

Next, the proposed method is evaluated on the intermediate datasets of the Tanks & Temples to validate the generalization ability without any model fine-tuning. The input image resolution is 1920 × 1056. Since the relative offset between multiple views on the Tanks & Temples datasets is smaller than that on the DTU datasets, we set the number of image views to 5. To fairly evaluate the generalization quality of the reconstructed point clouds, F-score is selected as the evaluation metric, which is the harmonic average of precision and recall under a given threshold $d$.


Fig. 7. Visualization of point cloud reconstruction results on Tanks & Temples datasets. It is clear that the proposed method can generate complete reconstruction with smooth details even on low-texture regions and large-scale scenes.

TABLE II
QUANTITATIVE F-SCORE COMPARISONS FOR RECONSTRUCTED POINT CLOUD ON TANKS AND TEMPLES DATASETS (HIGHER F-SCORE MEANS BETTER)

| Method | Mean | M60 | Train | Horse | Lighthouse | Family | Panther | Playground | Francis |
|---|---|---|---|---|---|---|---|---|---|
| COLMAP[6] | 42.14 | 44.83 | 42.04 | 25.63 | 56.43 | 50.41 | 46.97 | 48.53 | 22.25 |
| MVSNet[15] | 43.48 | 55.99 | 28.55 | 25.07 | 50.09 | 53.96 | 50.86 | 47.90 | 34.69 |
| Fast-MVSNet[40] | 47.39 | 49.16 | 42.91 | 34.98 | 47.81 | 65.18 | 46.2 | 53.27 | 39.59 |
| Point-MVSNet[11] | 48.27 | 51.97 | 43.06 | 34.20 | 50.79 | 61.79 | 50.85 | 52.38 | 41.15 |
| R-MVSNet[9] | 48.40 | 69.96 | 46.65 | 32.59 | 42.95 | 51.88 | 48.80 | 52.00 | 42.38 |
| P-MVSNet[22] | 55.62 | 55.08 | 54.29 | 40.22 | 65.20 | 70.04 | 55.17 | 60.37 | 44.64 |
| Ours | **56.05** | **59.89** | 50.01 | **47.64** | 51.47 | **75.52** | **55.23** | 57.17 | **51.52** |

Quantitative comparison results with other methods are illustrated in Table II. Point cloud reconstructed by our method can achieve the average F-score value of 55.75, which is better than other state-of-the-arts works. In particular, our method can obtain higher scores in all 8 testing scenarios than MVSNet and Point-MVSNet. Fig. 7 shows the quantitative point clouds of our method. It can be clearly seen that our network can reconstruct the complete scene and maintain smoother and finer details even in low-textured areas.

### *E. Ablation Studies*

To validate the effectiveness of the proposed model variants, a comprehensive ablation study is further conducted with detailed qualitative and quantitative comparisons on the DTU testing set. For quantitative comparisons, the average absolute depth error metric and prediction accuracy metric within a fixed distance threshold (2mm and 4mm are listed) are selected to evaluate the quality of the estimated depth map.

*1) Cost Volume Regularization Module Variants:* For the ablation study of the proposed cost regularization module, the visual comparisons are first conducted on scan23 and scan24 scenes. Note that for a fair comparison, the parameters of model training are all consistent. As shown in Fig. 8, benefited from our proposed global 3D convolution block and boundary refinement block, Fig. 8(e) can produce a clearer depth map in the object boundary than in Fig. 8(c). In addition, error maps in Fig. 8(d) and Fig. 8(f) represent the absolute difference between the predicted depth map and the ground truth, which means the brighter the pixel value, the greater the error. It can be clearly seen that compared with Fig. 8(d), the overall error generated in Fig. 8(f) is lower by adding the proposed cost volume regularization module. This might be attributed to the global 3D convolution with large and valid receptive fields, which can capture more geometric properties and effectively alleviate the incorrect feature matching correspondence in low-textured regions. Table III shows the quantitative results compared with the public-provided pre-trained model by MVSNet. As it can be seen, by introducing the proposed cost volume regularization module, the depth map predicted by MVSNet and our model reduces the average depth error and distance error.

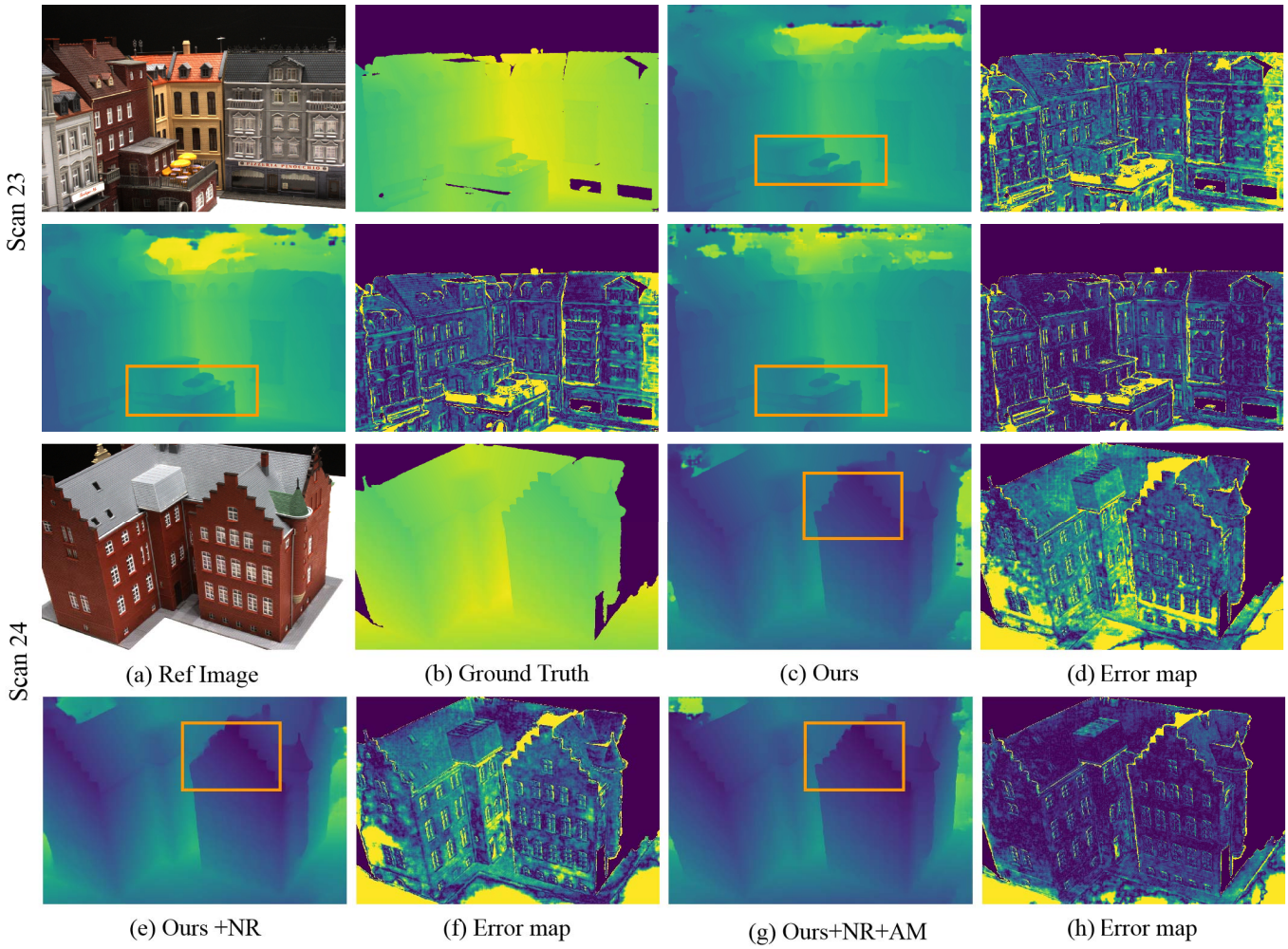



Fig. 8. Qualitative analysis for the proposed cost regularization module and cost aggregation mechanism. "Ours + NR" represents that we apply a new regularization module, which is consisted of a global convolution block and boundary refinement block. "Ours + AM" represents the new cost aggregation mechanism based on neighboring pixel-wise visibility. Error map denotes the absolute difference between the predicted depth value and ground truth (brighter areas indicate larger errors).

*2) Visibility-Aware Aggregation Module Variants:* Fig. 8 also shows the qualitative results of the proposed cost aggregation module variants. It can be seen that by combining the new adaptive aggregation mechanism to learn the pixel-wise visibility, the brightness of the error map in Fig. 8(h) is apparently lower than that in Fig. 8(f), and the overall quality of depth estimation is better even in low-textured regions (*e.g.*, the red brick walls of buildings). In addition, compared with our proposed cost regularization module, our AM can significantly reduce the depth estimation error and distance error, which indicates that high-quality cost aggregation is more effective than using well-designed cost regularization manners. In Table IV, an ablation study is conducted about the usage of AM on DTU datasets. By adding our proposed cost aggregation module, MVSNet can obtain a more precise depth map in terms of mean absolute depth error and distance error, especially the depth error is reduced by 17.8%. Compared with directly constructing aggregated spatial cost by AANet, the performance of cost aggregation for depth inference can be further improved by introducing the smoothness constraint to the pixel-visibility map. In addition, by integrating the proposed NR and AM, the mean absolute depth error and distance error are further declined. Table V shows the quantitative comparisons of our multi-stage depth inference model. Given the limited memory consumption and running time, the overall quality of depth inference improves remarkably as the number of stages increases.

TABLE III
THE STATISTICAL EVALUATION COMPARISON RESULTS OF DIFFERENT COST REGULARIZATION MODULES ON DTU DATASETS. THE NR DENOTES THE PROPOSED COST VOLUME REGULARIZATION MODULE

| Method | Resolution | Mean abs depth error | < 2mm | < 4mm |
|---|---|---|---|---|
| MVSNet | 1/4 × 1/4 | 11.63 | 63.13% | 79.95% |
| MVSNet + NR | 1/4 × 1/4 | 9.84 | 64.87% | 80.01% |
| Ours | 1 × 1 | 7.65 | 77.42% | 86.71% |
| Ours +NR | 1 × 1 | 6.27 | 80.99% | 89.18% |

*3) Loss Functions:* Quantitative results of the training model with the proposed loss functions are reported in Table VI. As can be seen that compared with $L_1$ loss function, directly applying $L_{uncertain}$ loss function can not achieve

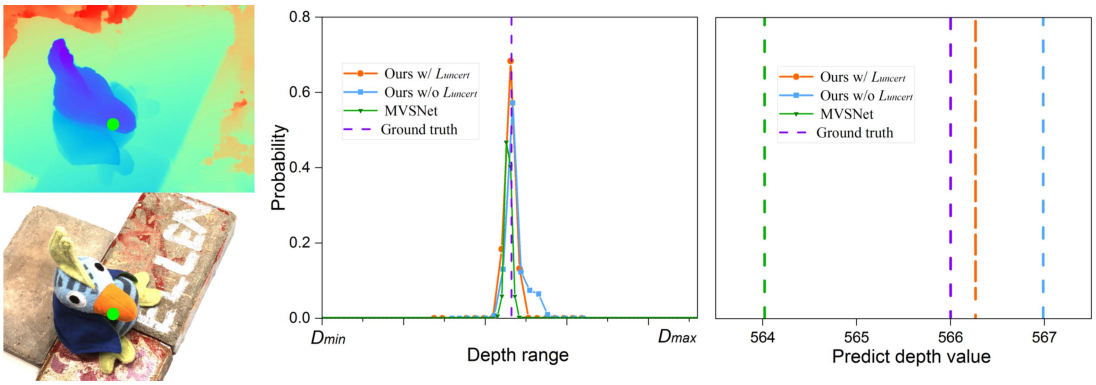



Fig. 9. Visualization of sampled point probability volume distribution on the DTU testing set. Left: sampled point on the reference image and predicted depth map. Middle: the probability distribution of the sampled point in MVSNet, our model with uncertain loss, and our model without uncertain loss. Right: the predicted depth value comparisons.

TABLE IV

THE STATISTICAL EVALUATION RESULTS OF DIFFERENT MODULES ON DTU DATASETS. THE AM DENOTES THE AGGREGATION MODULE OF COST VOLUME

| Method | Resolution | Mean abs depth error | < 2mm | < 4mm |
|---|---|---|---|---|
| MVSNet | 1/4 × 1/4 | 11.63 | 63.14% | 79.95% |
| MVSNet + AM (AANet) | 1/4 × 1/4 | 9.73 | 66.32% | 80.05% |
| MVSNet + our AM | 1/4 × 1/4 | 9.55 | 67.14% | 81.03% |
| Ours | 1 × 1 | 7.65 | 77.41% | 86.71% |
| Ours + AM | 1 × 1 | 6.11 | 83.47% | 89.86% |
| Ours + NR + AM | 1 × 1 | 6.03 | 83.94% | 90.16% |

TABLE V

THE STATISTICAL EVALUATION COMPARISON OF OUR MULTI-STAGE DEPTH ESTIMATION MODEL

| Method | Resolution | Mean abs depth error | < 2mm | < 4mm |
|---|---|---|---|---|
| Our 1st stage | 1/4 × 1/4 | 8.27 | 71.21% | 85.39% |
| Our 2nd stage | 1/2 × 1/2 | 6.39 | 79.72% | 88.53% |
| Our final model | 1 × 1 | 6.03 | 83.94% | 90.16% |

the desired depth estimation, which implies that $L_{uncertain}$ loss function is hard to balance the matching quality of probability volume and the error of depth estimation regression simultaneously. By further combing $L_1$ loss and $L_{uncertain}$ loss, our model can obtain superior performance in terms of the distance error and mean absolute depth error. This might be because $L_1$ loss function can easily regress the depth estimation error and $L_{uncertain}$ loss function can effectively improve the matching quality of the probability volume, which can further promote the depth regression precision through backpropagation without increasing the computation burden during the training.

*4) Visualization of the Probability Volume Distribution:* To comprehensively evaluate the matching quality of probability volume, a point is randomly sampled in low-textured regions on the DTU testing set and the difference in the predicted probability distribution with MVSNet is visualized. Note that the number of depth hypotheses in MVSNet is 192, while the number of our final stage is 8 and the number of depth

TABLE VI

EVALUATION OF THE PROPOSED LOSS FUNCTION

| $L_{uncertain}$ | $L_1$ | Mean abs. depth error | Prediction prec. (2mm) | Prediction prec. (4mm) |
|---|---|---|---|---|
| √ | | 8.27 | 71.21% | 85.39% |
| | √ | 6.19 | 82.83% | 89.46% |
| √ | √ | 6.03 | 83.94% | 90.16% |

hypotheses in the second stage is 16. Therefore, the predicted probability volume at the second stage is utilized to achieve better visualization. As shown in Fig. 9, compared with MVSNet, the peak of our estimated probability volume is distributed near the ground truth and the prediction result in the second stage is already close to the ground truth due to our coarse-to-fine manner and adaptive cost volume aggregation module. Besides, after adding the uncertainty loss function $L_{uncertain}$, the peak of the probability distribution is higher than the single $L_1$ loss, which further improves the accuracy of the estimated depth.

## V. CONCLUSION

In this paper, a multi-stage pixel-visibility learning-based stereo network is proposed. On the one hand, different from existing learning-based depth estimation networks that equally treat the contribution of each pixel to the cost volume aggregation and use the variance-based metric to filter the cost volume, the visibility of neighboring pixels is integrated and a self-adaptive cost volume aggregation is applied, which enables the network to learn the inherent constraints from matching correspondences in low-textured regions. Experimental results on some benchmarks show that the proposed aggregation mechanism can effectively aggregate cost volume in low-textured regions. On the other hand, the global convolution block and boundary refinement block are introduced instead of directly using the general 3D convolution to suppress incorrect feature matching for cost regularization. In addition, the uncertainty of predicted probability distribution is leveraged to design a new loss function, which aims to improve the reliability of depth estimation without increasing the parameters burden of the learned model. Comprehensive experimental comparisons with other methods are conducted

on the DTU and the Tanks & Temples datasets. The result shows that our method can achieve high-precision and high-density 3D scene reconstruction with real-time and limited memory consumption. In future work, we would like to remove the restriction of additional labeling of the camera parameters and achieve efficient unsupervised learning in the end-to-end multi-view depth estimation network.

## REFERENCES


[1] J. Zhang, Z. Liu, Y. Gao, and G. Zhang, "Robust method for measuring the position and orientation of drogue based on stereo vision," *IEEE Trans. Ind. Electron.*, vol. 68, no. 5, pp. 4298–4308, May 2021.

[2] A. Hussain, A. R. Memon, H. Wang, Y. Wang, Y. Miao, and X. Zhang, "S-VIT: Stereo visual-inertial tracking of lower limb for physiotherapy rehabilitation in context of comprehensive evaluation of SLAM systems," *IEEE Trans. Autom. Sci. Eng.*, vol. 18, no. 4, pp. 1550–1562, Oct. 2021.

[3] B. Lu *et al.*, "Toward image-guided automated suture grasping under complex environments: A learning-enabled and optimization-based holistic framework," *IEEE Trans. Automat. Sci. Eng.*, early access, Dec. 29, 2022, doi: 10.1109/TASE.2021.3136185.

[4] N. D. F. Campbell, G. Vogiatzis, C. Hernández, and R. Cipolla, "Using multiple hypotheses to improve depth-maps for multi-view stereo," in *Proc. Eur. Conf. Comput. Vis.*, Aug. 2008, pp. 766–779.

[5] S. Galliani, K. Lasinger, and K. Schindler, "Massively parallel multiview stereopsis by surface normal diffusion," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Jun. 2015, pp. 873–881.

[6] J. L. Schönberger et al, "Pixelwise view selection for unstructured multi-view stereo," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 501–518.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] S. Cheng *et al.*, "Deep stereo using adaptive thin volume representation with uncertainty awareness," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 2521–2531.

[13] X. D. Gu, Z. W. Fan, Z. Z. Dai, S. Y. Zhu, F. T. Tan, and P. Tan, "Cascade cost volume for high-resolution multi-view stereo and stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 2492–2501.

[14] J. Y. Yang, W. Mao, J. M. Alvarez, and M. M. Liu, "Cost volume pyramid based depth inference for multi-view stereo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 4876–4885.

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

[16] A. Yu *et al.*, "Attention aware cost volume pyramid based multi-view stereo network for 3D reconstruction," *ISPRS J. Photogramm. Remote Sens.*, vol. 175, pp. 448–460, May 2021.

[17] K. T. Giang, S. Song, and S. Jo, "Curvature-guided dynamic scale networks for multi-view Stereo," in *Proc. 10th Int. Conf. Learn. Represent.*, Apr. 2022, pp. 1–19.

[18] 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)*, Jun. 2019, pp. 4384–4393.

[19] 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.

[20] 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.

[21] R. T. Collins, "A space-sweep approach to true multi-image matching," in *Proc. CVPR IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Dec. 1996, pp. 358–363.

[22] K. Luo, T. Guan, L. Ju, H. Huang, and Y. Luo, "P-MVSNet: Learning patch-wise matching confidence aggregation for multi-view stereo," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10451–10460.

[23] E. Zheng, E. Dunn, V. Jojic, and J. M. Frahm, "Patchmatch based joint view selection and depthmap estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1510–1517.

[24] A. Kar, C. Häne, and J. Malik, "Learning a multi-view stereo machine," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 365–376.

[25] Z. Y. Wu, X. Y. Wu, X. P. Zhang, S. Wang, and L. L. Ju, "Semantic stereo matching with pyramid cost volumes," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 7483–7492.

[26] F. Zhang, V. Prisacariu, R. Yang, and P. H. S. Torr, "GA-Net: Guided aggregation net for end-to-end stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 185–194.

[27] X. Guo, K. Yang, W. Yang, X. Wang, and H. Li, "Group-wise correlation stereo network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3268–3277.

[28] S. H. Im, H. G. Jeon, and S. Lin, "DPSNet: End-to-end deep plane sweep stereo," in *Proc. 7th Int. Conf. Learn. Represent.*, Mar. 2019, pp. 1–12.

[29] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[30] H. Yi *et al.*, "Pyramid multi-view stereo net with self-adaptive view aggregation," in *Proc. Eur. Conf. Comput. Vis.*, Aug. 2020, pp. 766–782.

[31] X. Z. Zhu, H. Hu, S. Lin, and J. F. Dai, "Deformable ConvNets V2: More deformable, better results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 9308–9316.

[32] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 472–480.

[33] B. Ummenhofer *et al.*, "DeMoN: Depth and motion network for learning monocular stereo," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5038–5047.

[34] 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.

[35] 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.

[36] Q. Xu and W. Tao, "PVSNet: Pixelwise visibility-aware multi-view stereo network," 2020, *arXiv:2007.07714*.

[37] H. F. Xu and J. Y. Zhang, "AANET: Adaptive aggregation network for efficient stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 1956–1965.

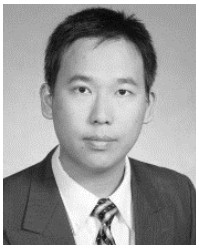
[38] F. Wang, S. Galliani, C. Vogel, P. Speciale, and M. Pollefeys, "PatchmatchNet: Learned multi-view patchmatch stereo," 2020, *arXiv:2012.01411*.

[39] 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.

[40] 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.


**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 Assistant 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 articles as the first and corresponding author in international journals and conferences. His research interests include exoskeleton robots, special equipment for mechatronics, and computer vision.

**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, China. His current research interests include 3D reconstruction, intelligent transportation systems, SLAM, and deep learning.



**Shan Jiang** received the Ph.D. degree in industrial and systems engineering from Rutgers University, New Brunswick, NJ, USA. He was a Research Assistant with the Center for Advanced Infrastructure and Transportation (CAIT), where he was involved in developing predictive models for driver risk analysis and using deep reinforcement learning for adaptive signal controls. He is currently the Data Science Lead with Johnson & Johnson Supply Chain, NJ, USA. His main research interests are modeling, optimization, and control of transportation systems and data-driven solutions for industrial applications using machine learning and data mining.

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

Also, he is currently 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, such as IEEE TRANSACTIONS ON ENGINEERING MANAGEMENT, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS, IEEE TRANSACTIONS ON INTELLIGENT VEHICLES, and IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING. His current research interests include intelligent transportation systems, neuroergonomics, brain and cognitive science, complex networks, non-parametric machine learning, reinforcement learning, and their applications in industrial or medical problems. He is a member of the IEEE Computational Intelligence Society and the Association for Computing Machinery. He has also been serving as a reviewer or a session chair for several top-tier international journals and conferences in his research field.

**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 a Professor with the Key Laboratory of System Control and Information Processing, Ministry of Education, Shanghai Jiao Tong University, China. He is also currently with the Science and Technology on Avionics Integration Laboratory, China Aeronautical Radio Electronics Research Institute, Shanghai, China. His research interests include deep learning, fatigue recognition, and human–machine interaction. He is currently 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.

**Guimin Chen** received the B.S., M.S., and Ph.D. degrees from the School of Mechatronics Engineering, Xidian University, Xi'an, China, in 2000, 2003, and 2005, respectively, all in mechanical engineering. He is currently a Full Professor with Xi'an Jiaotong University, Xi'an, and the Dean of the Institute of Intelligent Robots. He is currently an Associate Editor of IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING.