Multi-metric Depth Pyramid and Bilateral Segmentation Based Depth Inference for Multi-View Stereo

1st Jiaming Ji
School Of Electrical Engineering And Automation
Xiamen University Of Technology
Nanchang, China
644745100@qq.com

2nd *LumeiSu**
School Of Electrical Engineering And Automation
Xiamen University Of Technology
Xiamen, China
sulumei@163.com

3rd Zhihao Huang
School Of Electrical Engineering And Automation
Xiamen University Of Technology
Xiamen, China
956719363@qq.com

4th Jiajun Wu
School Of Electrical Engineering And Automation
Xiamen University Of Technology
Xiamen, China
1332451126@qq.com

2022 China Automation Congress (CAC) | 978-1-6654-6533-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/CAC57257.2022.10055806

***Abstract*—An efficient and efficient Pyramid Multiview Stereo (MVS) network, which is a 3D Bilateral Network (3D Bisenet), is presented to reconstruct the dense point cloud accurately and completely. First, a multimetric depth map pyramid is introduced, which can be used to reconstruct the shallow missing features by efficiently aggregating the depth estimates at the different layers. Then, a 3-D bipartite segmentation network is presented to keep matching pairs in all views and to enhance the detail properties of the model, so as to recover the lost scene. Furthermore, the high resolution and precision of the scene model can be obtained by combining the rapid down-sampling and the semiu-Net-type context paths. Experiments show that the proposed method achieves an excellent performance on DTU data sets, and significantly improves performance.**

***Index Terms*—3D Point Cloud Reconstruction, PBS-MVSNet, Multi-metric Depth Pyramid, Bilateral Segmentation**

I. Introduction

The goal of Multi-View Stereo (MVS) is to estimate its geometric representation from multiple calibrated images, and recover a dense 3D representation of scenes using stereo correspondences as the primary cue [1]. Traditional MVS methods[2] mostly use manually set matching variables to measure the consistency between multiple views, whereas there are more and more deep learning-based methods emerging as a result of extensive research and application of deep learning. These methods show improved completeness and granularity on many MVS benchmarks. However, some issues should be addressed in order to further improve the performance of reconstruction results.

Firstly, it is difficult to use feature extraction with 2D CNN such as MVSNet[3] and R-MVSNet[4] to dealing with thin or textureless structures,when the receptive field is in a fixed regular pixel grid, such as MVSNet[3] and R-MVSNet[4].

Secondly, few works in the field of cost volume aggregation focus on the issue of pixel-level visibility, and most methods apply U-Net structure[3]. The backbone network will lose a lot of contextual feature information that cannot be recovered simply by fusing shallow features, and the context-percetion function in the network is not well used for changes in texture richness in different regions, resulting in a reduction in final reconstruction quality.

In this paper, we propose a novel depth-based MVS method briefly named PBS-MVSNet that generates pyramid depth maps by building multi-scale pyramids and parallelizing them using BS-MVSNet. We design a 3D bilateral segmentation network in the cost volume aggregation part by analyzing the improved optimization of U-Net in the segmentation domain[16]. At the same time, according to the ACMM concept[35], multi-metric information cannot effectively improve the overall network's integrity. To iteratively obtain the optimal depth map, we use a multi-scale constraint to aggregate multi-metric depth information. As a result, the network proposed in this paper can generate a high-quality dense point cloud while obtaining a more accurate depth map and improving the processing speed of a single frame image. The following are our primary contributions:

(1)We build a multi-scale depth pyramid network to fuse multiple layers of depth information.Through the use of the shallow reliable depth map, the network can correct mismatched regions on the deep depth map, improving the accuracy and robustness of 3D reconstruction

(2) We propose a bilateral segmentation network regularization module to reduce the loss of shallow feature information.Furthermore, context feature extraction is performed using

This work is financially supported by Natural Science Foundation of department of science and technology of Fujian Province(No.2022J011255)

a fast CNN network,which reduces computation and memory consumption.

II. Related work

A. Traditional MVS

Traditional multi-view reconstruction methods often make use of the projection relation of multi-view to optimize the depth, which can be classified as volume approach [8], point based [6], and depth map fusion algorithm [7]. Depth map fusion is the easiest and the most adaptable. Schonberger et al [9] proposed an algorithm for reconstructing COLMAP. During the feature matching phase, they got hand-crafted features, optimized the depth of each pixel, and obtained a good effect. COLMAP is widely used in traditional MVS approaches because of its excellent performance in many situations. It is noted that there are drawbacks to depth map-based methods. All of these approaches depend on a predefined standard pixel level view selection, which is not applicable to different scenes.

B. Deep Learning MVS

Some problems can not be solved with traditional MVS methods, for example, the inability of learning and prediction of occluded objects. Recent advances in deep learning have helped to solve these problems. new concepts In 2018, Yao et al proposed MVSNet [3] approach, which improves the performance of MVS. In this paper, a new method was proposed to extract the features and construct the homographic matrix transformation from a reference image and multiple source images. However, during the iteration of the algorithm, its cost and memory consumption have a cubic relation. It needs a lot of memory and computation, but it also provides a big frame and improved thinking for the future rebuilding of the network. In order to solve this problem, Yao et al. suggested using Gated Recurrent Unit (GRU) instead of MVSNet's 3D-CNN [4] to minimize memory loss in reasoning phase. On the basis of this, Zizhuang Wei and his colleagues proposed a hybrid RNN-CNN model to standardize the cost, and a method of inter-view cost-volume aggregation in order to increase computing efficiency and precision. On the other hand, CasMvsnet [12] offers an alternative MVSNet solution to reduce the amount of running memory.

We propose an adaptie cost volume regularzation module based on the appeal analysis that fuses deatiled information without losing spatial information to improve deep feature matching. To fuse depth information at multiple scales, we build a multi-metric depth graph pyramid network.Through breadth and geometric consistency,the network suppresses mismatch errors and obtains more reliable depth information without increasing the amount of computation.

III. Method

A. Overall Framework

As shown in Figure 1, the overall architecture of the proposed PBS-MVSNet follows the model of learning based MVS pipelines. In Section B.After that, we present the details of Section C.Moveover.. In section D, we propose a 3D bipartite cost volume regularization network. Finally, the depth map regression and the training loss are introduced in Section E and Section F respectively.

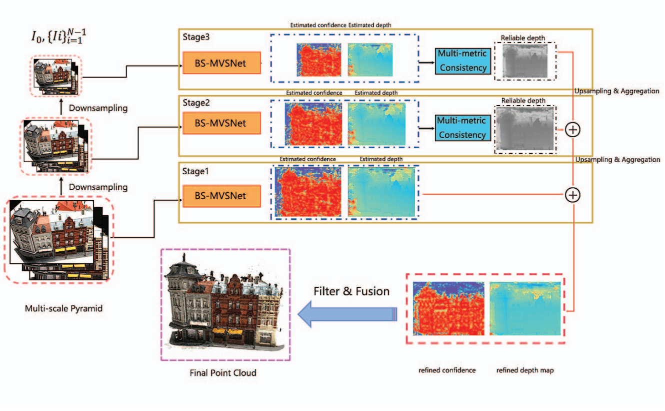

Fig. 1. Overall framework of the proposed method

Regarding the overall procss framework, we initially estimate the depth map of each reference image $I_i = 0$in sequence. The weights of the 2D CNN feature extraction network are shared among the multi-view images $I_i = 0...N-1$shared between multiple views, where N is the number of multiple views. Then, on the estimated depth map, the final 3D point cloud model is obtained by filtering and merging. In this paper, we construct a multiscale depth pyramid with k (k = 0, 1, 2, 3) to improve the precision and robustness of reconstruction. Then we run the images of each layer using BS-MVSNet and regress the desired confidence to produce depth maps at different scales. Then, the depth map of each layer is output to the next level, and the depth map is refined from the coarse to the fine, so that the mismatch error between the image and the image is corrected, and then a more precise depth map can be obtained.

B. Multiscale Depth Pyramid

In order to solve the problem of matching blur, we employ a multiscale deep pyramid aggregation method, which utilizes light and geometry consistency. Then, we use a coarse-to-fine structure to model the mismatch.

Firstly, geometric and photometricconsistency is required for accurate, reliable, and well-matched depths obtained on low-level stage k. Then, in the pyramid, we gradually replace the corresponding upper-level high-resolution blurred depth $P^{k+1}(p)$ with the downsampling layer $k+1$ low-level reliable depth $P^k(p)$, where $P^k(p)$ represents the confidence map at pixel point p. We use the camera parameter $K_i(p)$ to project the uv coordinates of the image I_i pixel point p onto the neighbor image I_j and convert them to neighborhood coordinates. To extract features, we define the pixel point p projected to the neighborhood as p_{source}.Then ,using the camera parameter of source images $K_j(p_{source})$,back-project p_{source} to the reference image to obtain the back-projected pixel point $p_{resource}$ and the corresponding camera parameters $K_j(p_{resource})$. We keep the pixel p that statisfy the geometric consistency which constraints should be met:

$$\|p - p_{resource}\|_2 < 1 \tag{1}$$

$$\|K_i(p) - K_j(p_{resource})\|_1 < 0.01 \cdot K_i(p) \tag{2}$$

At last, we can obtain a more accurate high-confidence depth map and reconstruct a higher-quality 3D point cloud by continuously propagating and replacing blurred depth at high resolution with low-scale accurate depth.

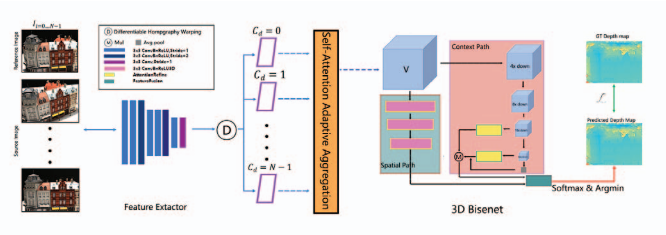

Fig. 2. Structure of the BS-MVSNet network

C. Adaptive Cost Volume Aggregation

As the BS-MVSNet shown in Fig.2. , we continue to use the CNN network for feature extraction.Then the output features will be transformed from the feature maps between different views to the reference camera's conical stereo space using the differential homography transformation to create its 3D planar scan feature volume $f_{i=0...N-1}$.

When constructing the cost volume, many of the previous methods assume that the contribution values of all views are the same. And these method will use variance-based mapping for statistical aggregation to re-weight the contributions of differents pixels to construct the required cost volume.

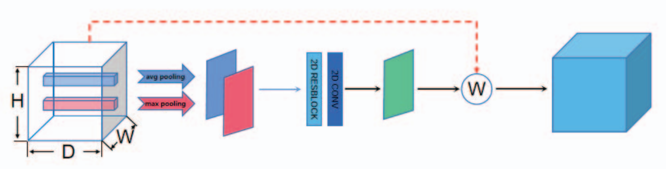

Fig. 3. Pixel-wise view adaptive aggregation module

However, in practice, multi-view images would have issues such as different lighting conditions, reflections, and so on. As a result, it is obviously unreasonable to expect all perspectives to contribute equally. As a result, in order to assess the difference between various points of view, the attention The mechanism's pixel-wise view [14] is depicted in Fig. 3, and the adaptive aggregation module of the voxel-wise view [15] is depicted in Fig. 4. A selective weighted self-attention mechanism is introduced in the height and width dimensions in the pixel-level view aggregation process to consider the shared focus weights under the depth assumption, and the cost volume $c_{d,h,w}$ is defined as:

$$f'_{i,d,h,w} = f_{i,d,h,w} - f_{0,d,h,w} \tag{3}$$

$$c_{d,h,w} = \frac{\sum_{i=1}^{N-1}(1 + w_{h,w}) \bigodot f'_{i,d,h,w}}{N-1} \tag{4}$$

Here $w_{h,w}$is a 2-dimensional weighted attention map that encodes the relationship between pixels in different views, and $bigodot$denotes element-wise multiply operation.

Every voxel has its own distinguishing feature in the voxel view, and it is necessary to assume different depth assumptions. So, by introducing a 3D weighted attention map to select useful cost information, the cost volume $c_{d,h,w}$is defined as:

$$c_{d,h,w} = \frac{\sum_{i=1}^{N-1}(1 + w_{d,h,w}) \bigodot f'_{i,d,h,w}}{N-1} \tag{5}$$

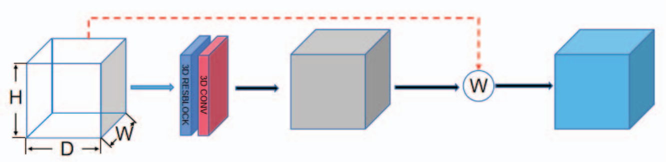

Fig. 4. Adaptive aggregation module at voxel-wise

D. Cost Volume Regularization

We design a 3D BiseNet[16] structure for cost volume regularization to effectively use local map information and multi-scale contextual information. As shown in Table I, it is a novel 3D semi-U-Net structure with contextual information path (CP) and spatial information path (SP) components.

Context Path : In order to obtain a large receptive field and improve computational efficiency, we use the Xception [19] framework in the backbone network to build a network that can rapidly downsample the features of the cost volume. The global context information is then obtained via average pooling, and the global context information and the features of the previous two stages of downsampling are finally fused.

Simultaneously, to extract more detailed features during this period, we used a semi-UNet structure in the last two stages of downsampling and introduced an attention mechanism refinement module (3D AttentionRefinementModule, 3D ARM), the structure of which is shown in Table I(a). The final output feature will be directly input into the feature vector's fusion module (3D FeatureFusionMoudle, 3D FFM), which can easily obtain rich context information without upsampling, which greatly reduce the calculation amount in the SP path and speed up the net computational efficiency.

Spatial Path : Spatial information can be obtained from the spatial encoding path. We use a spatial encoding path with three layers, each layer contains a three-dimensional

TABLE I
DETAILED STRUCTURES OF ATTENTION REFINEMENT AND FEATURE FUSION MODULES

Input	Layer	Output
	3DAttentionRefinementModule	
$c_{d,w,h}$	ConvBnReLU,kernel=3 $\times$ 3,stride = 1	fb_0
fb_0	Conv3D,kernel=1,stride = 1	fb_1
fb_1	BatchNorml	fb_2
fb_2	Sigmod	fb_3
fb_3	mul	$f_{c_{d,h,w}}$
	(a)Attention refinement module structure	
	3D FeatureFusionModule	
$[f_{c_{d,h,w}}, f_{s_{d,h,w}}]$	concate,dim=1	$fb3d\_0$
$fb3d\_0$	ConvBnReLU3D,kernel=1 $\times$ 1,stride = 1	$fb3d\_1$
$fb3d\_1$	Mean,dim=(2,3)(global pool)	$fa3d\_0$
$fa3d\_0$	Conv3D,kernel=1,stride=1	$fa3d\_1$
$fa3d\_1$	BatchNorml	$fa3d\_2$
$fa3d\_2$	Conv3D,kernel=1,stride=1	$fa3d\_3$
$fa3d\_3$	Sigmod	$fa3d\_4$
$[fb3d\_1, fa3d\_4]$	mul	$fb3d\_2$
$[fb3d\_1, fb3d\_2]$	add	$f_{costvolum}$
	(b)Feature fusion module structure	

convolution with kernel_size = 3, stride =2, followed by 3D BatchNormal and ReLU modules for batch processing. The final feature map is the source image $\frac{1}{2^n}$ (n is the number of spatial encoding path layers), that is $\frac{1}{8}$. Since the original cost volume has a high resolution,we can extract a wealth of spatial information from it.

$\mathbf{Feature\ Fusion\ Module(FFM)}$: The context path produces shallow contextual semantic information, whereas the spatial path produces deep spatial detail information. To merge two moudles, We build a feature fusion module that combines the two outputs.The concept is based on SENet [20], and the detailed structure is shown in Table I. (b).We reclassify the combined features into a feature vector by learning an attention weight which can effectively suppress unnecessary features and obtain a more accurate feature map.

E. Depth map regression

In order to get the normalized cost, we need to do a deep regression to get the depth map and the confidence map, and then we use the softmax algorithm on the depth dimension to return to the original size of the image. Pixel depth regression and reliability measure of estimate

Simultaneously, we estimate the regression depth value D(p) expectation using argmin on the probability volume P in order to generate continuous depth estimations:

$$D(p) = \sum_{j=0}^{D-1} d_j \cdot P(p, j) \tag{6}$$

where $P(p, j)$ represents the confidence of all pixels in the depth hypothesis d_j.

F. Loss function

A supervised learning strategy applied in the network. To guide the depth estimation, we use the real ground depth as the network's supervision signal. The loss is estimated by the average absolute difference between the ground depth and the estimated depth as Loss, just like MVSNet[3]. The loss function for each sample is:

$$\mathcal{L} = \sum_{x \in x_{valid}} \|d_j(x) - \hat{d_j}(x)\|_1 \tag{7}$$

where $x_{valid}, d_j(x)$ and $\hat{d_j}(x)$ respectively represent the set of effective feature pixel in the ground truth, the estimated depth map and the ground truth.

IV. EXPERIMENT

A. Implementation Details

$\mathbf{Training}$:Our PBS-MVSNet is based on a DTU data set [23]. In this paper, we train the network on the training data set and evaluate it on the evaluation data set. The image size in training is WXH = 1600 × 1200, and the simultaneous input of multiple views is N = 4. The depth sampling hypothesis is set to [450 mm, 950 mm], and the planar sweep depth is 192. In this paper, we proposed a deep neural network based on Pytorch [24] and trained it over 12 epochs using Adam [25], with an initial learning rate of 0.001 and a decline of 0.9. The training is carried out on one Nvidia RTX2080Ti graphics card with a batch size of 1.

TABLE II
QUANTITATIVE FINDINGS FROM THE DTU EVALUATION SET (LOWER IS BETTER).

Method	Acc.(mm)	Comp.(mm)	overall(mm)
Camp[27]	0.835	0.554	0.695
Colmap[28]	0.400	0.664	0.532
Furu[29]	0.613	0.941	0.777
Gipuma[31]	**0.283**	0.873	0.578
Tola[30]	0.342	1.190	0.766
SurfaceNet[10]	0.450	1.040	0.745
MVSNet[3]	0.396	0.527	0.462
R-MVSNet[4]	0.406	0.434	0.420
PointMVSNet[25]	0.361	0.421	0.391
P-MVSNet[26]	0.406	0.434	0.420
CIDER[32]	0.417	0.437	0.427
Fast-MVSNet[34]	0.336	0.403	0.370
D2HC-MVSNet[5]	0.395	0.378	0.386
Vis-MVSNet[33]	0.369	0.361	0.365
PVA-MVSet[15]	0.378	**0.336**	0.357
Ours	0.335	0.367	**0.351**

$\mathbf{Evaluation}\ mathbfTest$:In the DTU data set, we have a data set of N = 4 images and a depth plane scan of D = 192.

$\mathbf{Filtering}\ mathbfa\ mathbfFusion$:All of the expected depth maps will be merged to form one point cloud [3] In our experiments, we only consider depth values with a reliable confidence of 0.8. Then, we use geometric consistency to select the pixels that are found in more than three adjacent views, and then project them into 3-D.

B. DTU dataset results

Performance Indicators : The evaluation indicators provided by the DTU dataset are followed.While Gipuma's method is currently the most accurate, we outperform it in terms of overall quality through Table II.

Our method reconstructs a more complete dense point cloud, and as shown in Table II, the PBS-MBSNet's speed of extracting a single frame image is significantly faster than several other methods, demonstrating the high speed and overall performance.

Runtime and Memory Performance : Table III. shows, where H, W, D represent the height, width and the number of depth sample respectively displays the time and memory performance of each method, which was tested on an RTX 2080Ti. It can be found that the network shows a considerable improvement in running time when compared to previous ways. The multi-scale depth pyramid can be processed independently in parallel.

TABLE III
SINGEL FRAME RUNTIME AND MEMORY CONSUMPTION

Method	H,W,D	times/per view	Mem.
MVSNet[3]	1600,1184,256	4.7s	15.4GB
RMVSNet[4]	1600,1184,512	9.1s	6.7GB
PatchmatchNet[13]	1600,1200,256	1.25s	**5.5GB**
PointMVSNet[25]	1280,960,96	1.69s	7.2GB
PVAMVSNet[15]	1600,1184,192	0.91s	10.2GB
Ours	1600,1200,192	**0.67s**	8.3GB

V. CONCLUTION

In this paper, we propose a multiview stereo vision network, which is based on a bipartite cost-volume-regularized network. The BS-MVSNet is designed to reduce the loss of shallow context information at the same time, and to reduce the occurrence of mismatching. The MLN will build a more precise depth map based on photometric and geometrical agreement, which will increase the precision and robustness of the whole network. Experiments indicate that PBS-MVSNet can obtain a higher total score on DTU data sets than other approaches.

REFERENCES

[1] Strecha, Christoph, et al. "On benchmarking camera calibration and multi-view stereo for high resolution imagery." 2008 IEEE conference on computer vision and pattern recognition. Ieee, 2008.

[2] Schönberger, Johannes L., et al. "Pixelwise view selection for unstructured multi-view stereo." European Conference on Computer Vision. Springer, Cham, 2016.

[3] Yao, Yao, et al. "Mvsnet: Depth inference for unstructured multi-view stereo." Proceedings of the European Conference on Computer Vision (ECCV). 2018.

[4] Yao, Yao, et al. "Recurrent mvsnet for high-resolution multi-view stereo depth inferen ce." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

[5] Yan, Jianfeng, et al. "Dense hybrid recurrent multi-view stereo net with dynamic consistency checking." European Conference on Computer Vision. Springer, Cham, 2020.

[6] Furukawa, Yasutaka, and Jean Ponce. "Accurate, dense, and robust multiview stereopsis." IEEE transactions on pattern analysis and machine intelligence 32.8 (2009): 1362-1376.

[7] Xu, Qingshan, and Wenbing Tao. "Multi-scale geometric consistency guided multi-view stereo." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

[8] Kutulakos, Kiriakos N., and Steven M. Seitz. "A theory of shape by space carving." International journal of computer vision 38.3 (2000): 199-218.

[9] Schonberger, Johannes L., and Jan-Michael Frahm. "Structure-from-motion revisited." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[10] Ji, Mengqi, et al. "Surfacenet: An end-to-end 3d neural network for multiview stereopsis." Proce edings of the IEEE International Conference on Computer Vision. 2017.

[11] Hartmann, Wilfried, et al. "Learned multi-patch similarity." Proceedings of the IEEE international conference on computer vision. 2017.

[12] Gu, Xiaodong, et al. "Cascade cost volume for high-resolution multi-view stereo and stereo matching." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[13] Wang, Fangjinhua, et al. "PatchmatchNet: Learned Multi-View Patchmatch Stereo." Proceed ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

[14] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.

[15] Yi, Hongwei, et al. "Pyramid multi-view stereo net with self-adaptive view aggregation." Europe an Conference on Computer Vision. Springer, Cham, 2020.

[16] Yu, Changqian, et al. "Bisenet: Bilateral segmentation network for real-time semantic segmentation." Proceedings of the European conference on computer vision (ECCV). 2018.

[17] Wei, Zizhuang, et al. "Aa-rmvsnet: Adaptive aggregation recurrent multi-view stereo network." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[18] Çiçek, Özgün, et al. "3D U-Net: learning dense volumetric segmentation from sparse annotation." International conference on medical image computing and computer-assisted intervention. Springer, Cham, 2016.

[19] Chollet, François. "Xception: Deep learning with depthwise separable convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[20] Rahimzadeh, Mohammad, and Abolfazl Attar. "A modified deep convolutional neural network for detect-

ing COVID-19 and pneumonia from chest X-ray images based on the concatenation of Xception and ResNet50V2." Informatics in medicine unlocked 19 (2020): 100360.

[21] Collins, Robert T. "A space-sweep approach to true multi-image matching." Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 1996.

[22] Aanæs, Henrik, et al. "Large-scale data for multiple-view stereopsis." International Journal of Computer Vision 120.2 (2016): 153-168.

[23] Paszke, Adam, et al. "Automatic differentiation in pytorch." (2017).

[24] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).

[25] Chen, Rui, et al. "Point-based multi-view stereo network." Proceedings of the IEEE/CVF Internati onal Conference on Computer Vision. 2019.

[26] Luo, Keyang, et al. "P-mvsnet: Learning patch-wise matching confidence aggregation for multi-view stereo." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.

[27] Campbell, Neill DF, et al. "Using multiple hypotheses to improve depth-maps for multi-view stereo." European Conference on Computer Vision. Springer, Berlin, Heidelberg, 2008.

[28] Schönberger, Johannes L., et al. "Pixelwise view selection for unstructured multi-view stereo." European Conference on Computer Vision. Springer, Cham, 2016.

[29] Furukawa, Yasutaka, and Jean Ponce. "Accurate, dense, and robust multiview stereopsis." IEEE transactions on pattern analysis and machine intelligence 32.8 (2009): 1362-1376.

[30] Galliani, Silvano, Katrin Lasinger, and Konrad Schindler. "Massively parallel multiview stereopsis by surface normal diffusion." Proceedings of the IEEE International Conference on Computer Vision. 2015.

[31] Tola, Engin, Christoph Strecha, and Pascal Fua. "Efficient large-scale multi-view stereo for ultra high-resolution image sets." Machine Vision and Applications 23.5 (2012): 903-920.

[32] Xu, Qingshan, and Wenbing Tao. "Learning inverse depth regression for multi-view stereo with correlation cost volume." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. No. 07. 2020.

[33] Jingyang Zhang, Yao Yao, Shiwei Li, Zixin Luo, and Tian Fang." Visibility-aware multi-view stereo network." British Machine Vision Conference (BMVC), 2020.1

[34] Yu, Zehao, and Shenghua Gao. "Fast-mvsnet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[35] Xu, Qingshan, and Wenbing Tao. "Multi-scale geometric consistency guided multi-view stereo."Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.