Hybrid Attention Cascade Multi-View Stereo Network

Weiqiang Liu[1], Rongshang Chen[1], Huarong Xu[1], Lifen Weng [2*]
[1] College of Computer and Information Engineering, Xiamen University of Technology, China
[2] School of Design and Art, Xiamen University of Technology, China
*2009990509@xmut.edu.cn

2022 IEEE 16th International Conference on Anti-counterfeiting, Security, and Identification (ASID) | 978-1-6654-9067-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ASID56930.2022.9995975

Abstract—**The multi-view stereo reconstruction method based on deep learning is usually affected by the weak-textured area or occlusion in the real scene. Therefore we propose a multi-view stereo reconstruction network method with a hybrid attention mechanism. A hybrid attention module is added to the feature extractor to improve the performance in weak-textured regions. In order to reduce occlusion effects a module is used to adjust the view weights. We find adding depth-adaptive partitioning will improve the performance of our method. Our method is trained and tested on the DTU and Tanks and Temples datasets, the results show that our method has good results in terms of reconstruction accuracy and completeness.**

Keywords—hybrid attention; multi-view stereo; depth estimation; 3d reconstruction

I. Introduction

The purpose of multi-view stereo methods is to obtain 3D models of real-world scenes from multiple images, which can be used in virtual reality, autonomous driving, and cultural relics protection. Traditional MVS methods[1-5] use hand-crafted image features matching metrics to measure multi-view consistency, thereby reconstructing scene geometry and corresponding texture information. Recent deep learning-based methods[6-9] have shown superior accuracy and higher completeness on many MVS benchmark sets[10-13]. Through the convolutional neural network, the effect of feature extraction and cost volume regularization is better. However, some challenging problems remain to be solved to further improve the reconstruction quality. The first is that it is difficult to extract effective image features for texture-less or weak-textured surfaces, and the second is the handling of occlusion and illumination changes in multiple view, which limit the robustness and completeness of 3D reconstruction.

Recent attempts based on MVSNet[8, 14-16] introduced a cascade mechanism that from coarse to fine to improve depth estimation but did not provide a perfect solution for problems such as occlusion and illumination from different perspectives. To solve the problems mentioned above, we propose a cascaded multi-view stereo reconstruction network based on hybrid attention. The main improvements of this method are as follows:

- We propose a hybrid attention multi-scale feature extraction module to improve the effect of the feature extraction module for weak-textured or texture-less areas.
- We propose a weight adjustment module for feature maps from different perspectives to deal with occlusion and illumination changes under different perspectives.
- We propose a depth adaptive segmentation module, which adaptively adjusts the depth segmentation according to the cost volume of each view and improves the accuracy of reconstruction.

II. Related Work

A. Multi-view Stereo Reconstruction Method

According to the output scene representation, traditional MVS reconstruction methods can be divided into three types, voxel-based[17, 18], point cloud-based[4, 19], and depth-based methods[1, 2, 5, 20]. Volumetric methods first discretize the entire 3D space into regular cubes and then use a photometric consistency measure to determine whether a voxel belongs to a surface. Methods to discretize space into voxels[21] are memory-intensive, so these methods are difficult to apply to large-scale scenes. Whereas point-based methods[22] focus on 3D points in space, usually starting from a sparse set of matching key points and progressively making progressively denser reconstructions using a propagation strategy. In contrast, depth-based methods exhibit greater flexibility in the 3D geometric modeling of the scene. It simplifies MVS reconstruction to a relatively simple problem of estimating depth maps for each view and can obtain point clouds or voxels by fusion. Many successful traditional MVS algorithms have been proposed to generate depth maps. COLMAP[5] uses hand-crafted features and jointly estimates pixel-level view selection, depth maps, and surface normal to exploit assumptions such as photometric and geometric priors. Although traditional MVS methods yield impressive results, they exploit hand-crafted image features with unreliable photometric consistency on non-Lambertian surfaces, low-texture, and texture-less regions.

B. Deep Learning-Based Multi-View Geometry

Recent research on MVS abandons the use of traditional hand-crafted image features and instead utilizes deep convolutional networks for better reconstruction accuracy and

978-1-6654-7735-2/22/$31.00 ©2022 IEEE

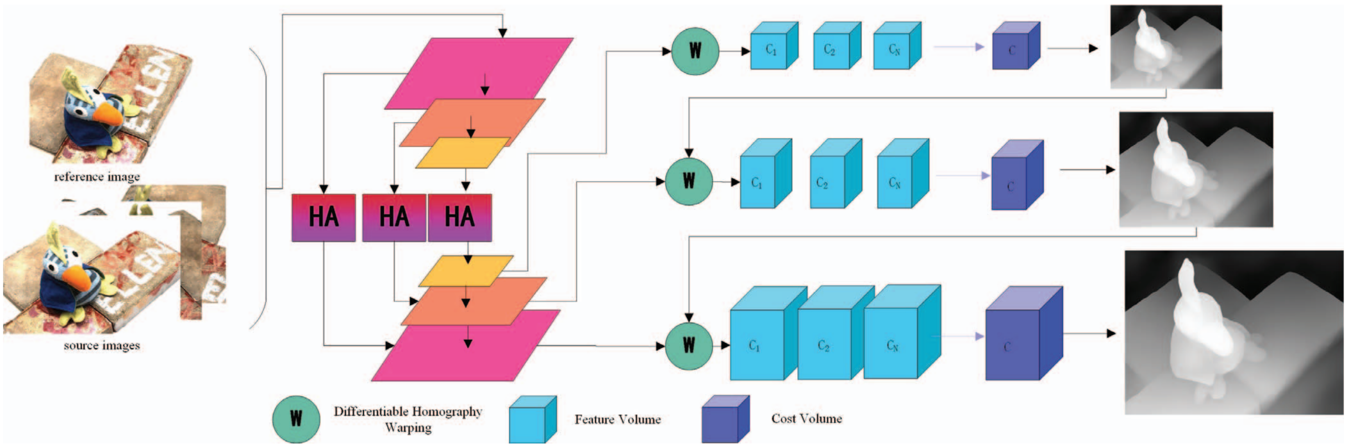

Figure 1. Network architecture of the proposed method.

completeness. Voxel-based methods such as SurfaceNet[23] and LSM[24] were first proposed, they construct cost volumes from multi-view images, and regularize and infer voxels through the cost volumes. However, SurfaceNet and LSM are limited to small-scale reconstructions due to the disadvantage of utilizing voxel representations. The depth-based MVSNet[6] takes a reference image and several other source images as input and extracts depth image features, and then constructs a 3D cost volume through differentiable homography transformation in the network. To reduce the huge memory consumption of MVSNet, recently proposed variants of MVSNet are presented, which can be divided into multi-stage methods and recurrent methods. Multi-stage methods, such as CasMVSNet[15], CVP-MVSNet[16], UCS-Net[25], use coarse-to-fine predictions of low-resolution depth maps with large depth intervals and iterative up-sampling and refinement with a narrow depth range. This strategy makes the network occupy less memory for running. Other recurrent network methods such as R-MVSNet[7] and D2HC-RMVSNet[8], which use a recurrent neural network to sequentially regularize the cost volume along the depth dimension, also reduce memory usage.

III. Method

The network structure of this paper is shown in Fig. 1. The overall architecture of HA-CasMVSNet in this paper follows the typical framework of learning-based MVS methods. The input is images from multiple view of the scene and corresponding camera parameters, wherein the input images are divided into 1 reference image and N-1 source images. For the input N images, its image features ($H*W*F$) are extracted by an encoder with shared weights, and a 3D cost volume is constructed by transforming the features of the source image into the reference image frustum through a differentiable homography, The 3D cost volume for each view match is computed by matching the features and depth assumptions of the N-1 transformed source and reference images, the pixel-level mapping between the reference image and the i^{th} source image and depth. The homography transformation formula of is

$$\mathbf{H}_i^{(d)} = d\mathbf{K}_i\mathbf{T}_i\mathbf{T}_{ref}^{-1}\mathbf{K}_{ref}^{-1} \tag{1}$$

where $\mathbf{K}$ and $\mathbf{T}$ denote camera intrinsics and extrinsics respectively. Then the matching cost volume for each view is calculated by:

$$\mathbf{c}_i^{(d)} = \left(\mathbf{f}_{src_i}^{(d)} - \mathbf{f}_{ref}\right)^2 \tag{2}$$

We use a view-based neural network to adjust the weights of all cost volumes, then merge them into the reference cost volume, and the probability volume is obtained through a depth-regularized 3D convolutional network. For each pixel, the depth is calculated by:

$$\mathbf{D} = \sum_{d=d_{\min}}^{d_{\max}} d \times \mathbf{P}(d) \tag{3}$$

where $\mathbf{P}(d)$ is the probability estimation for all pixels at depth d.

Finally, a dense point cloud of the scene is obtained by filtering and fusing the depth maps of all images.

A. Feature Extraction Module Using Hybrid Attention

As mentioned earlier, the 3D cost volume is constructed by matching the 2D feature maps extracted from different viewpoints, so obtaining discriminative and reliable features is important in the MVS method. In 3D reconstruction, it is generally believed that reflective surfaces and low-texture or texture-less regions are the main difficulties for ordinary CNNs to extract feature parts. In these challenging regions such as lacking texture, features at different scales and regions with different texture richness are adaptively aggregated through attention aggregation.

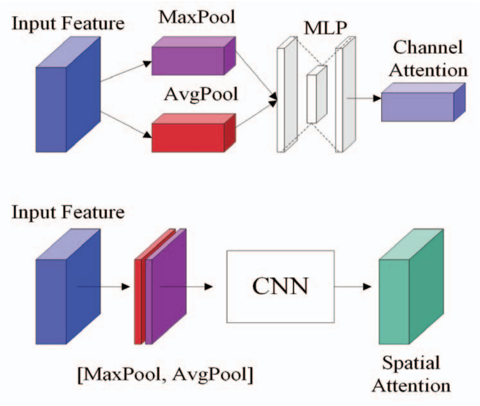

Figure 2. Channel and spatial Attention

We apply channel and spatial attention modules on the skip connections of the feature extraction network to compute complementary attention, and the specific network structure is shown in Fig. 2. For the spatial attention part, we adopt a pooling operation to aggregate the channel information of the feature map and then generate the required 2D channel attention map through a convolutional network. In short, the spatial attention calculation is

$$M_S(F) = \sigma(f([\text{AvgPool}(F); \text{MaxPool}(F)])) \quad (4)$$

where σ represents the sigmoid function, and f represents the convolution operation.

For channel attention, we use the relationship between feature channels to generate channel attention maps. For the spatial information of aggregated feature maps, we generate spatial context descriptors through pooling operations and then generate corresponding channels through a multi-layer perceptron. The calculation formula of channel attention is

$$M_C(F) = \sigma(MLP(\text{AvgPool}(F)) + MLP(\text{MaxPool}(F))) \quad (5)$$

where σ represents the sigmoid function, where the weights of the MLP are shared between the two inputs.

B. Weight Adjustment Module for Different Perspectives

After constructing the previous cost volumes, the next step is to aggregate all cost volumes into one cost volume one for regularization. A common practice is to average feature volumes, with the consideration that all views should be considered equally important. However, considering that different shooting angles may lead to problems such as occlusion and changes in lighting conditions on non-Lambertian surfaces, depth estimation is more difficult.

Therefore, as illustrated in Fig. 3, we design a weight adjustment module to handle unreliable matching costs, which is defined as

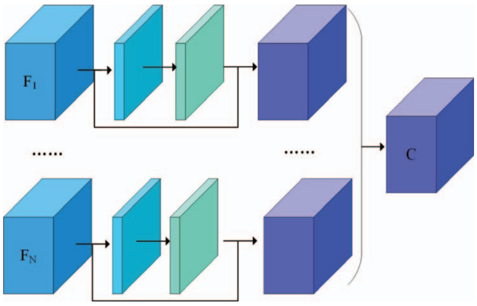

Figure 3. Cost volume weight adjustment module

$$\mathbf{C}^{(d)} = \frac{1}{N-1}\sum_{i=1}^{N-1}\left[1+\omega\left(F_i^{(d)}\right)\right]\odot F_i^{(d)} \quad (6)$$

where $\odot$ represents Hadamard multiplication, and ω is a pixel-by-pixel weight adjustment adaptively generated according to the feature volume of each view so that the key contextual information is enhanced.

C. Adaptive Depth Partitioning Module

Rather than using uniform depth sampling division in the CasMVSNet[15] method, we propose an adaptive depth division strategy to discretize the depth interval D into N intervals, as shown in Fig. 4.

We predict the final depth as a linear combination of the centers of each bin, so that each stage can output a relatively smoothly varying depth map. We use a convolutional network plus a multi-layer perceptron to estimate the width ratio of each bin,

$$b_i = \frac{b_i' + \epsilon}{\sum_{j=1}^{N}\left(b_j' + \epsilon\right)} \quad (7)$$

Where $\epsilon =$ 10-3, such a relatively small positive number can ensure that the width of each bin is strictly positive.

D. Loss Function

In this paper, the intermediate output and prediction results of the output of the different stages of the cascaded cost volume, we supervise all the outputs, and the total loss is defined as

$$\text{Loss} = \sum_{k=1}^{N} \lambda^k \cdot L^k \quad (8)$$

where λ^k refers to loss weight and L^k refers to the loss at the k^{th} stage.

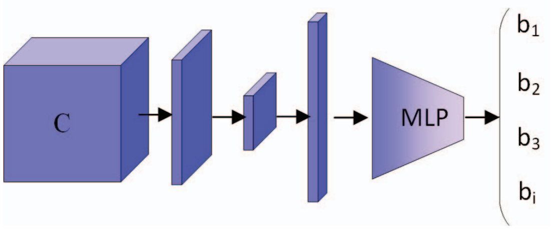

Figure 4. Adaptive depth partitioning module

IV. Experiments

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

A. Datasets and Implementation Details

1) Datasets DTU is a large-scale MVS aspect dataset consisting of 124 different scenes scanned from multiple viewing angles under a variety of different lighting conditions. The Tanks and Temples dataset contains multiple real scenes with a small depth range. Specifically, it includes scenes such as Family, Francis, Horse, Lighthouse, M60, Panther, Playground, and Train. We train our method on the DTU training set and test it on the DTU evaluation set. To verify the generalizability of our method, we also test it on tanks and temples, using models trained from DTU and not fine-tuned on this dataset.

2) Implementation Details We train our HA-CasMVSNet on a DTU dataset consisting of 79 different scenarios. We used the training set of the previous MVS method, we resized the original image resolution to 640x512 for training, the number of input images was set to N=5, the number of depth assumptions was D=192, and adaptive sampling was performed from 425mm to 935mm. We implemented and trained our network via PyTorch. We train for 16 epochs with an initial learning rate of 0.001, which is halved after 10, 12, and 14 epochs.

B. Experiment Results

Showing the quantitative results on the DTU validation dataset in Table I, where we use the official MATLAB evaluation code to calculate accuracy and completeness. Compared with other method, our method outperforms other methods and has some improvement in completeness. The qualitative results on DTU dataset are shown in Fig. 5, and we can see that our method generates more complete and finer point cloud details. In addition, our method tests our trained model on the Tanks and Temples dataset, and the corresponding quantitative results are shown in Table II. The qualitative point cloud results of the intermediate set of Tanks and Temples benchmark are visualized in Fig. 6. Our method significantly improves reconstruction quality in some scenes.

TABLE I. Quantitative Results on DTU Evaluation Dataset

Method	Acc.(mm)	Comp.(mm)	Overall(mm)
Gipuma[2]	0.283	0.873	0.578
COLMAP[5]	0.400	0.664	0.532
R-MVSNet[7]	0.385	0.459	0.422
D2HC-RMVSNet[8]	0.395	0.378	0.386
CasMVSNet[15]	0.325	0.385	0.355
CVP-MVSNet[16]	0.296	0.406	0.351
EPP-MVSNet[26]	0.413	0.296	0.355
HA-CasMVSNet(Ours)	0.332	0.358	0.345

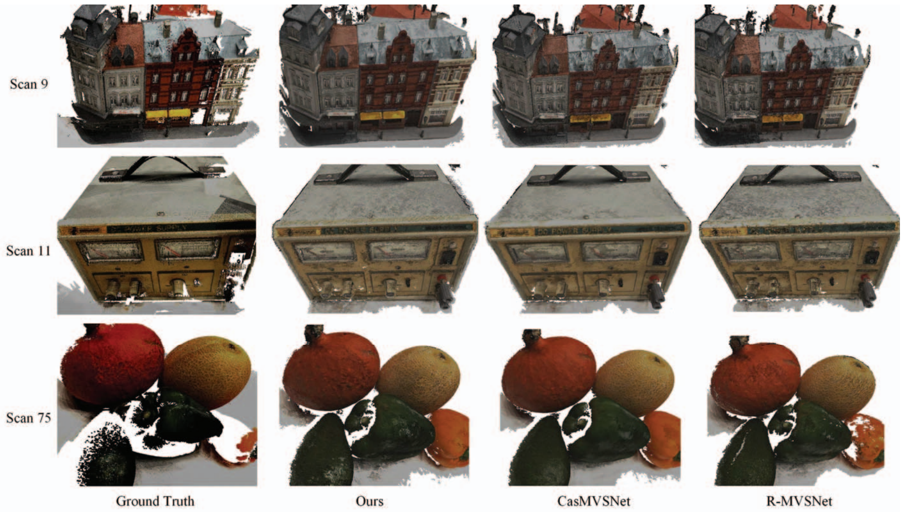

Figure 5. Qualitative comparisons of Scan9, 11 and 75 in DTU dataset

TABLE II. QUANTITATIVE RESULTS ON DTU EVALUATION DATASET

Method	mean	Fam.	Fran.	Horse	Light.	M60	Pan.	Play.	Train
COLMAP[5]	42.14	50.41	22.25	25.63	56.43	44.83	46.97	48.53	42.04
R-MVSNet[7]	48.40	69.96	46.65	32.59	42.95	51.88	48.80	52.00	42.38
D2HC-RMVSNet[8]	59.2	74.69	56.04	49.42	60.08	59.81	59.61	60.04	53.92
CasMVSNet[15]	56.42	76.36	58.45	46.20	55.53	56.11	54.02	58.17	46.56
CVP-MVSNet[16]	54.03	76.5	47.74	36.34	55.12	57.28	54.28	57.43	47.54
EPP-MVSNet[26]	61.68	77.86	60.54	52.96	62.33	61.69	60.34	62.44	55.3
HA-CasMVSNet(Ours)	57.79	77.92	61.83	48.94	57.64	60.06	53.82	56.42	45.72

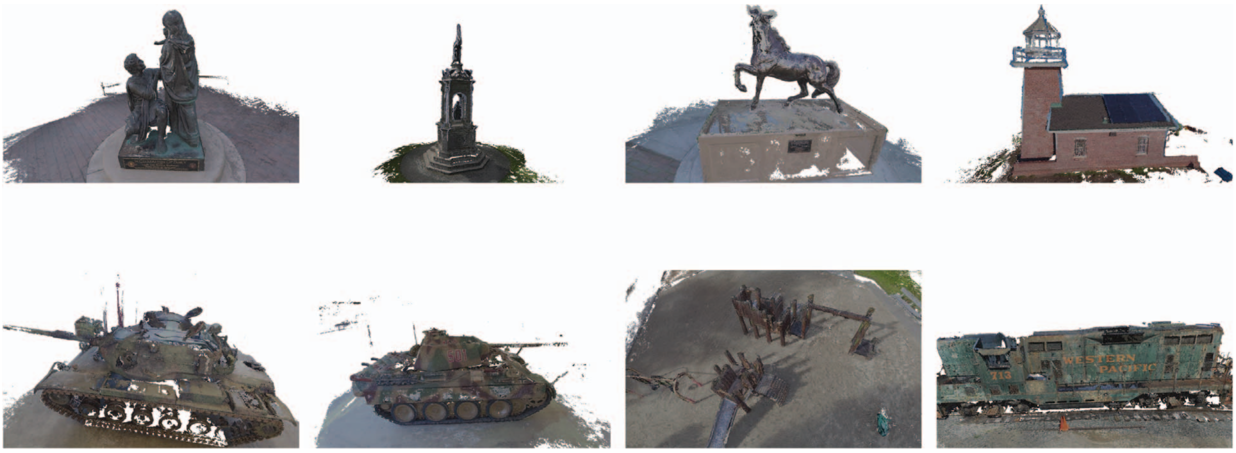

Figure 6. Point cloud results on the intermediate set of Tanks and Temples dataset

C. Ablation Experiment

In this section, we provide ablation experiments to quantitatively analyze the effectiveness of each method. The following ablation studies are performed on the DTU dataset with the same parameters as in Section 4.2. We compare with the baseline model by reverting the corresponding changes. The results of the reconstruction accuracy and completeness of different components during the training process are shown in the table III. It can be seen that each independent improvement can reduce the overall prediction error, and the complete model can achieve the best performance.

TABLE III. QUANTITATIVE PERFORMANCE WITH DIFFERENT COMPONENTS ON DTU EVALUATION DATASET

Model	Acc.(mm)	Comp.(mm)	Overall(mm)
Baseline	0.325	0.385	0.355
+HA-FeatureExtraction	0.343	0.361	0.352
+weight adjustment	0.324	0.374	0.349
+adptive depth-sampling	0.314	0.392	0.353
Full	0.332	0.358	0.345

V. CONCLUSION

We propose a novel multi-stage multi-view stereo network based on mixed attention, called HA-CasMVSNet. For each view feature extraction, the effect of weak-textured surfaces is effectively improved by mixing spatial and channel attention-aware features. The depth partitioning module further improves the performance of the model in testing. Our method shows competitive results on the DTU dataset, while the results on the Tanks and Temples dataset show that our method has good generality and scalability. However, due to the addition of attention mechanism, the parameters are increased and memory consumption is larger. An interesting future direction is to improve the reconstruction accuracy while reduce the memory consumption.

ACKNOWLEDGMENT

This paper was supported by the National Natural Science Foundation of China (No.U1805264, No.61772444); Fuxiaquan National Independent Innovation Demonstration Zone collaborative innovation platform project (3502ZCQXT2021009)； Fujian science and technology plan projects (2022Y0077)

REFERENCES

[1] N. D. Campbell, G. Vogiatzis, C. Hernández and R. Cipolla, "Using multiple hypotheses to improve depth-maps for multi-view stereo," in European Conference on Computer Vision: Springer, pp. 766-779, 2008.

[2] S. Galliani, K. Lasinger and K. Schindler, "Massively parallel multiview stereopsis by surface normal diffusion," in Proceedings of the IEEE International Conference on Computer Vision, pp. 873-881, 2015.

[3] C. Barnes, E. Shechtman, A. Finkelstein and D. B. Goldman, "PatchMatch: A randomized correspondence algorithm for structural image editing," ACM Trans. Graph., vol. 28, no. 3, p. 24, 2009.

[4] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," IEEE transactions on pattern analysis and machine intelligence, vol. 32, no. 8, pp, 1362-1376, 2009.

[5] J. L. Schönberger, E. Zheng, J.-M. Frahm and M. Pollefeys, "Pixelwise view selection for unstructured multi-view stereo," in European conference on computer vision: Springer, pp. 501-518, 2016.

[6] Y. Yao, Z. Luo, S. Li, T. Fang and L. Quan, "Mvsnet: Depth inference for unstructured multi-view stereo," in Proceedings of the European conference on computer vision (ECCV), pp. 767-783, 2018.

[7] Y. Yao, Z. Luo, S. Li, T. Shen, T. Fang, and L. Quan, "Recurrent mvsnet for high-resolution multi-view stereo depth inference," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 5525-5534, 2019.

[8] J. Yan, Z. Wei, H. Yi, M. Ding, R. Zhang, Y. Chen et al., "Dense hybrid recurrent multi-view stereo net with dynamic consistency checking," in European conference on computer vision: Springer, pp. 674-689, 2020.

[9] K. Luo, T. Guan, L. Ju, Y. Wang, Z. Chen, and Y. Luo, "Attention-aware multi-view stereo," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1590-1599, 2020.

[10] H. Aanæs, R. R. Jensen, G. Vogiatzis, E. Tola and A. B. Dahl, "Large-scale data for multiple-view stereopsis," International Journal of Computer Vision, vol. 120, no. 2, pp, 153-168, 2016.

[11] A. Knapitsch, J. Park, Q.-Y. Zhou and V. Koltun, "Tanks and temples: Benchmarking large-scale scene reconstruction," ACM Transactions on Graphics (ToG), vol. 36, no. 4, pp, 1-13, 2017.

[12] T. Schops, J. L. Schonberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys et al., "A multi-view stereo benchmark with high-resolution images and multi-camera videos," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3260-3269, 2017.

[13] Y. Yao, Z. Luo, S. Li, J. Zhang, Y. Ren, L. Zhou et al., "Blendedmvs: A large-scale dataset for generalized multi-view stereo networks," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1790-1799, 2020.

[14] H. Yi, Z. Wei, M. Ding, R. Zhang, Y. Chen, G. Wang et al., "Pyramid multi-view stereo net with self-adaptive view aggregation," in European Conference on Computer Vision: Springer, pp. 766-782, 2020.

[15] 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 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2495-2504, 2020.

[16] J. Yang, W. Mao, J. M. Alvarez and M. Liu, "Cost volume pyramid based depth inference for multi-view stereo," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4877-4886, 2020.

[17] K. N. Kutulakos and S. M. Seitz, "A theory of shape by space carving," International journal of computer vision, vol. 38, no. 3, pp, 199-218, 2000.

[18] S. M. Seitz and C. R. Dyer, "Photorealistic scene reconstruction by voxel coloring," International Journal of Computer Vision, vol. 35, no. 2, pp, 151-173, 1999.

[19] M. Lhuillier and L. Quan, "A quasi-dense approach to surface reconstruction from uncalibrated images," IEEE transactions on pattern analysis and machine intelligence, vol. 27, no. 3, pp, 418-433, 2005.

[20] Q. Xu and W. Tao, "Multi-scale geometric consistency guided multi-view stereo," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5483-5492, 2019.

[21] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison et al., "Kinectfusion: Real-time dense surface mapping and tracking," in 2011 10th IEEE international symposium on mixed and augmented reality: Ieee, pp. 127-136, 2011.

[22] P. Merrell, A. Akbarzadeh, L. Wang, P. Mordohai, J.-M. Frahm, R. Yang et al., "Real-time visibility-based fusion of depth maps," in 2007 IEEE 11th International Conference on Computer Vision: Ieee, pp. 1-8, 2007.

[23] M. Ji, J. Gall, H. Zheng, Y. Liu and L. Fang, "Surfacenet: An end-to-end 3d neural network for multiview stereopsis," in Proceedings of the IEEE International Conference on Computer Vision, pp. 2307-2315, 2017.

[24] A. Kar, C. Häne and J. Malik, "Learning a multi-view stereo machine," Advances in neural information processing systems, vol. 30, 2017.

[25] S. Cheng, Z. Xu, S. Zhu, Z. Li, L. E. Li, R. Ramamoorthi et al., "Deep stereo using adaptive thin volume representation with uncertainty awareness," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2524-2534, 2020.

[26] X. Ma, Y. Gong, Q. Wang, J. Huang, L. Chen, and F. Yu, "EPP-MVSNet: Epipolar-assembling based Depth Prediction for Multi-view Stereo," in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5732-5740, 2021.