N2MVSNET: NON-LOCAL NEIGHBORS AWARE MULTI-VIEW STEREO NETWORK

Zhe Zhang[1], *Huachen Gao*[1], *Yuxi Hu*[2], *Ronggang Wang*[1*]

[1]School of Electronic and Computer Engineering, Peking University, Shenzhen, China
[2]School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, China

ABSTRACT

Learning-based multi-view stereo (MVS) methods have been widely studied recently. However, current works are limited to using fixed-size convolution kernels, leading to sub-optimal features which lack anisotropy in low-textured regions and tend to produce invalid depth blending at the edge of the foreground and background. In this paper, we propose N2MVSNet, which learns adaptive non-local neighbors matching (ANNM) and their spatial impact to overcome these deficiencies. Furthermore, we explore the ability of spatial perception to depth dimension and propose the 3D ANNM. Besides, following the coarse-to-fine scheme, severe mismatchings in coarser stages will result in error accumulation and propagation in finer stages. To this end, we adopt the pre-trained RGB guided depth refinement for depth hypothesis repolish. The robustness of the training process is further elevated by the energy aggregation loss. Extensive experiments on the DTU and Tanks and Temples datasets demonstrate that the proposed network achieves state-of-the-art results.

***Index Terms*—** Multi-view stereo, 3D reconstruction, depth estimation, RGB guided depth refinement

1. INTRODUCTION

Multi-view Stereo (MVS) plays a significant role in the representation and comprehension of 3D scenes, which is widely applied in many areas, e.g., autonomous driving, virtual reality, robotics, etc. Given a series of unstructured calibrated images, MVS methods construct the dense 3D point clouds by building the corresponding matches along an approximate depth range. Current methods can be categorized into direct point cloud-based [1], volumetric-based [2], and depth map-based [3], where the last representation decouples the challenging problem into per-view depth estimation and multi-view depth fusion, having been widely adopted recently.

Traditional MVS methods [4] usually follow a four-step pipeline, including random initialization, propagation, multi-view matching cost evaluation, and refinement. However, they often suffer from low-texture, exposure, and illumination changes. With the rapid growth of deep learning technologies, learning-based MVS methods have shown remarkable progress. In particular, Yao *et al.* [5] build the cost volume via differentiable homography warping and using 3D CNN [6] for robust cost matches. R-MVSNet [7] adopts GRUs for sequential regularization to reduce memory consumption. CasMVSNet [8], UCS-Net [9], and CVP-MVSNet [10] estimate depth maps in a coarse-to-fine manner. However, these methods use convolutions with fixed-size kernels, which lack anisotropy in low-textured regions and produce false depth blending at the edges of the foreground and background. AA-RMVSNet [11] imports deformable convolution to extract image features, and PatchmatchNet [12] augments the traditional propagation and cost evaluation steps of Patchmatch with learnable modules. However, these methods do not fully exploit the pixel-wise correlation between neighbors.

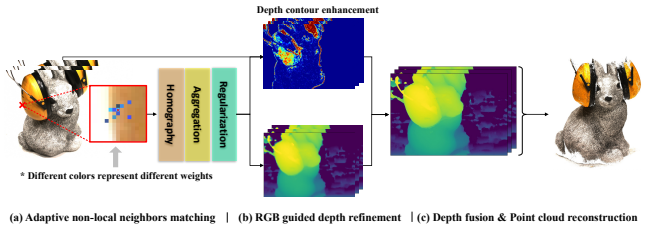

Fig. 1. Visualization of the brief multi-view stereo pipeline and the effectiveness of proposed modules.

Our key idea is based on the observation that the depth of a pixel is strongly related to neighbors around it. For example, neighbors inside the edge have a positive effect on a foreground pixel, while the background neighbors have the opposite impact (Fig. 1(a)). Therefore, we propose the Adaptive Non-local Neighbors Matching (ANNM) strategy, to find a set of sampled neighbors in a relatively large region and their spatial impact adaptively. More specifically, two neural sub-modules are used for producing spatial similarity weights and sampling offsets respectively, supplying the non-local neighbors aware feature extraction. Furthermore, we extend the ability of spatial perception to depth dimension and propose the 3D ANNM. Besides, following the cascade MVS framework, finer stages predictions cannot be corrected when coarser stages produce severe misprediction, which can

*Ronggang Wang is the corresponding author (rgwang@pkusz.edu.cn).

This work is supported by the National Natural Science Foundation of China U21B2012 and 62072013, Shenzhen Cultivation of Excellent Scientific and Technological Innovation Talents RCJC20200714114435057, Shenzhen Research Projects of 201806080921419290.

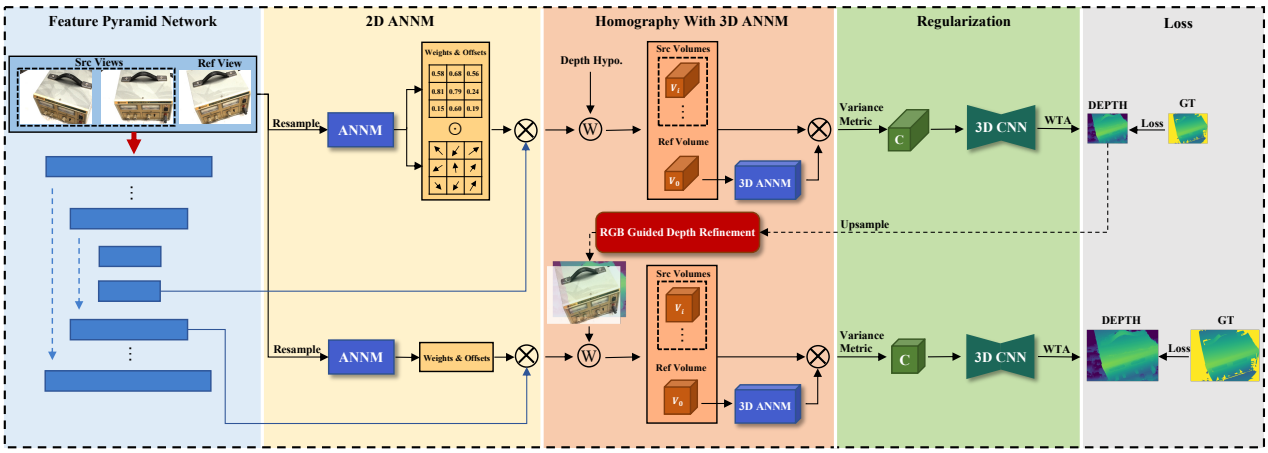

Fig. 2. **Network structure.** We adopt the coarse-to-fine scheme and two stages are presented for demonstration. Sec. 2.2 will introduce the adaptive non-local neighbors matching (ANNM) and its 3D extension, RGB guided depth refinement will be presented in Sec. 2.3, and the energy aggregation loss is shown in Sec. 2.4.

lead to the accumulation and propagation of errors. We use a pre-trained RGB guided depth refinement network to repolish the coarsest depth estimation and highlight the contours (Fig. 1(b)). In addition, a newly designed energy aggregation loss is proposed for better convergence.

In summary, the primary contributions are threefold:

- We propose the Adaptive Non-local Neighbors Matching (ANNM) strategy, which leverages the pixel-wise spatial correlation of neighbors and extends it as the voxel-wise 3D ANNM for preferable depth perception.
- We apply RGB guided depth refinement to repolish mispredictions in coarser stages by highlighting the contours valuable, and preventing the accumulation and propagation of errors for finer stages.
- The proposed network is extensively evaluated on the DTU dataset [13] and Tanks and Temples dataset [14]. The proposed method achieves state-of-the-art performance from extensive experimental results.

2. METHODOLOGY

2.1. Overall Pipeline

Given a set of unstructured images $\{I_i\}_{i=0}^N$, the reference image is I_0 and the other source images are $\{I_i\}_{i=1}^N$. Our proposed method predicts the depth map D aligned with I_0. The feature maps $\{F_i\}_{i=0}^N$ are extracted firstly by a feature pyramid network (FPN) [15] with shared weights. Afterward, a set of feature volumes $\{V_i\}_{i=0}^N$ is obtained via differentiable homography warping w.r.t. depth hypothesis d which uniformly sampled from $[d_{min}, d_{max}]$ to warp the 2D features into the 3D space in the reference camera frustum. Then, multiple feature volumes are aggregated into a single cost volume C to handle an arbitrary number of viewpoints. The 3D-CNN [6] is used for cost regularization to construct the probability volume P along the depth direction. Finally, the depth map D is generated from P by applying the winner-takes-all [16].

We adopt the coarse-to-fine structure [8–10] where the depths produced in coarser stages are used as the guidance of the depth hypothesis in finer stages and the exquisite details will be gradually restored. In the following sections, we consider the kth-level of the cascade structure for simplicity.

2.2. Adaptive Non-local Neighbors Matching

Denote one pixel $p(x, y)$ on image I, the proposed Adaptive Non-local Neighbors Matching (ANNM) aims to build a non-local neighbors aware feature for p by finding a set of sampled pixels in a relatively large region and their spatial correlation adaptively. First, as shown in Fig. 3(a), a channel-wise unification and a resample operation are applied to I. Inspired by [17], the blue channel has stronger robustness, we combine the RGB channels ($I_{R,G,B}$) to a unified representation as:

$$\hat{I}_{uni} = w_R \times I_R + w_G \times I_G + w_B \times I_B, \tag{1}$$

where $w_R = 0.1, w_G = 0.3$, and $w_B = 1.0$. Then the $\hat{I}_{uni}$ is downsampled by $stride = r$, repeated and stacked to get the resampled image $\hat{I}$ in the shape of $r^2 \times H/r \times W/r$. The receptive field becomes more effective after resampling, and weights are shared across different channels.

After the image transform, two sub-modules, $\mathcal{K}$ and $\mathcal{G}$, of ANNM are proposed for producing non-local spatial similarity weights W^p of $k^2 \times 1 \times 1$ and sampling offsets O^p of $2k^2 \times 1 \times 1$ for $p(x, y)$ respectively, where k is the size of the weighted kernel. The specific expressions are as follows:

$$\begin{cases} W^p = Sigmoid(\mathcal{K}(\{\frac{x}{r}, \frac{y}{r}\} \leftarrow \hat{I})) \\ O^p = \mathcal{G}(\{\frac{x}{r}, \frac{y}{r}\} \leftarrow \hat{I}) \end{cases}. \tag{2}$$

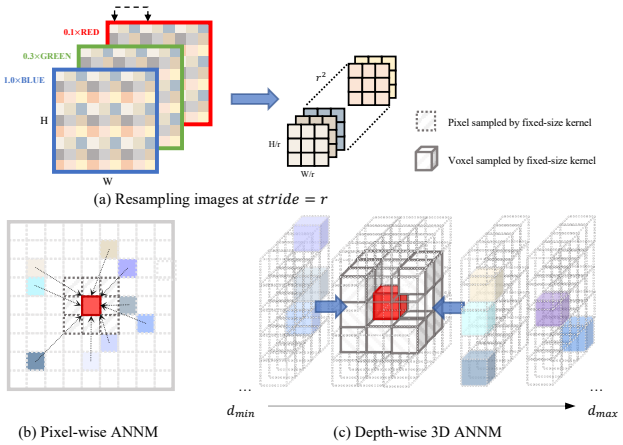

Fig. 3. **Illustrations of ANNM module**. (a) The channel-wise unification and resample. (b) The matching among the center pixel and colored (represent different weights) neighbors. (c) The matching among the center voxel and neighbors.

As shown in Fig. 3(b), each pixel in neighbors aware feature is obtained by calculating the weighted average from the learned weights and offsets. Denote F_{base} the feature extracted from classical FPN (local fixed-size kernel), the final non-local neighbors aware feature F can be computed as:

$$F = F_{base} \otimes ps(\frac{1}{N}\sum_{p\in\hat{I}} W^p \odot (O^p_{\Delta x} \cup O^p_{\Delta y})), \quad (3)$$

where $O_{\Delta x}$ denotes the horizontal offsets while $O_{\Delta y}$ is the vertical one, $\odot$ denotes element-wise multiplication and $N = k^2$ is the number of sampled pixels around center pixel. We apply pixel-shuffle (ps) [18] to transform the non-local similarity from the feature domain to the spatial domain and form the shape of the reference image.

Sec 2.1 mentioned that 2D features are encoded into the 3D camera frustum. Then, an intuitive idea is that we further extend the 2D ANNM to 3D dimension for better exploring explicit non-local depth aware cost matches. The 3D ANNM is applied to the reference feature volume separately for the simple reason that our ultimate goal is estimating the depth map towards I_0. As shown in Fig. 3(c), denote V_{base} the feature volume directly warped from reference feature F_0, the two 3D sub-modules learn the non-local depth-wise similarity weights W^q of $k^3 \times 1 \times 1$ and 3D sampling offsets O^q of $3k^3 \times 1 \times 1$. The final depth aware feature volume V integrates depth perception from voxels around, computing as:

$$V = V_{base} \otimes \frac{1}{N'}\sum_{q\in V_{base}} W^q \odot (O^q_{\Delta x} \cup O^q_{\Delta y} \cup O^q_{\Delta d}), \quad (4)$$

where q is the voxel in V_{base}, number of sampled voxels $N^{'} = k^3$, and $O_{\Delta d}$ denotes depth-wise offsets. The learned depth aware feature volume indicates more robust neighbors matching, and $\{V_i\}_{i=0}^N$ from all viewpoints are then warped into single cost volume to regress the depth estimation.

2.3. RGB Guided Depth Refinement

Depth maps generated from coarser stages are used as guidance for finer layers. However, severe incorrect prediction at coarser stages will be accumulated in finer stages. To this end, a recent SOTA RGB guided depth refinement network DCT-Net [19] is applied in the coarsest layer to better highlight the contours valuable for depth refinement. We upsample the coarse depth to the shape of I_0 which meets the high-res RGB contour information aided. Denote Φ^D and Φ^{I_0} are the features of D and I_0 generated from the feature extraction module, while the edge attention weights $\mathcal{W}^{I_0}$ learned from I_0 used to prevent texture over-transfer between the RGB/depth image. The refined depth $\hat{D}$ can be obtained as:

$$\hat{D} = \mathcal{DCT}(\Phi^D, \Phi^{I_0}, \mathcal{W}^{I_0}). \quad (5)$$

In detail, $\mathcal{DCT}(\cdot,\cdot,\cdot)$ can be computed as follows:

$$\Phi^E \triangleq \hat{\lambda}\mathcal{L}^2\left(\Phi^{I_0}\right) \odot \mathcal{W}^{I_0} + \Phi^D, \quad (6)$$

$$\hat{D} = \mathcal{F}^{-1}\left\{\mathcal{F}\left(\Phi^E\right) \oslash \left(I + \hat{\lambda}K^2\right)\right\}, \quad (7)$$

where $\hat{\lambda}$ is set as a learnable parameter, $\mathcal{L}$ denotes the Laplacian filter, $\mathcal{F}$ is the DCT operation while $\mathcal{F}^{-1}$ is the inverse one, I is the identity matrix and $\oslash$ denotes element-wise division, K is the 2D basis image of I_0. The outliers predicted by the coarser stages will be corrected through refinement.

2.4. Loss Function

We formulate the MVS problem in a classification manner [7, 20, 21] under the supervision of the ground-truth distribution P_{GT} among the set of valid pixels Ω. And $Loss_{CE}$, the cross-entropy loss between the predicted depth representation P and P_{GT}, is first utilized for constraint. Moreover, we exploit the property of DCT transform with de-correlation and energy concentration [22] to propose the innovative energy aggregation restriction $Loss_{EA}$ as:

$$Loss_{EA} = -\sum_{p\in\Omega} P_{GT}(p)\log\left\{\mathcal{F}^{-1}[\mathcal{F}(P(p))]\right\}. \quad (8)$$

The total loss of our l-stages model is formulated as:

$$Loss = \sum_{i=1}^{l}(\lambda_1 Loss^i_{CE} + \lambda_2 Loss^i_{EA}), \quad (9)$$

where $\lambda_1 = \lambda_2 = 0.5$ in our experiments.

3. EXPERIMENTS

3.1. Datasets

Our experiment is conducted on DTU [13], Tanks and Temples [14] and BlendedMVS [23] datasets. DTU is an indoor dataset consisting of 124 objects captured from 49 viewpoints with 7 different lighting conditions, and it contains the

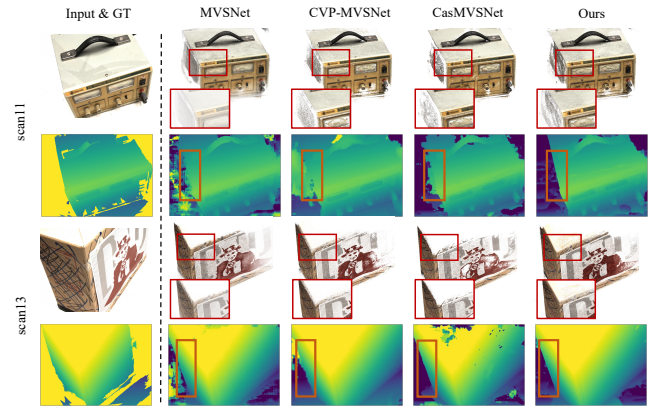

Fig. 4. **Qualitative results of scan 11 and scan 13 on DTU evaluation set[13].** The first and the third row are generated point clouds while the rest are the estimated depth maps.

Fig. 5. **Qualitative results on Tanks and Temples dataset.**

ground-truth point clouds for evaluation. Our training, validation, and evaluation split are the same as defined in [5] and [8] for a fair comparison. Tanks and Temples dataset provides a more challenging realistic environments with large-scale variations. BlendedMVS is a recently published large-scale synthetic dataset, consisting of over 17000 rendered images.

3.2. Results on DTU Dataset

Following the common practice, the number of views is set to 5 with a resolution of 640×512, the initial depth planes hypothesis is set to 48, and sampling interval is from 425*mm* to 935*mm*. Our model is implemented using PyTorch, trained for 16 epochs on NVIDIA Tesla V100 GPU with a starting learning rate of 0.001. For evaluation, we input 7 images with original resolution, and follow previous methods [5] for depth fusion to reconstruct point clouds. Qualitative results are shown in Fig. 4 and Tab. 1 shows quantitative results. We report accuracy (Acc.) and completeness (Comp.) statistics by the official MATLAB evaluation protocol, and our method achieves the best performance compared to other methods.

3.3. Results on Tanks and Temples Dataset

To further demonstrate the generalization ability of our method, the network is further trained on BlendedMVS dataset with 7 input images of 786×576 resolution, and evaluated on the Tanks and Temples benchmark. The F-score

Table 1. **Quantitative results on the DTU[13] and Tanks and Temples[14] datasets. Bold** numbers represent the best and underlined numbers represent the second-best.

Method	DTU			Tanks and Temples								
	Acc.	Comp.	Overall.↓	Family	Francis	Horse	LH	M60	Panther	PG	Train	Mean↑
COLMAP[24]	0.400	0.664	0.532	50.41	22.25	25.63	56.43	44.83	46.97	48.53	42.04	42.14
MVSNet[5]	0.396	0.527	0.462	55.99	28.55	25.07	50.79	53.96	50.86	47.90	34.69	43.48
R-MVSNet[7]	0.383	0.452	0.417	69.96	46.65	32.59	42.95	51.88	48.80	52.00	42.38	48.40
CasMVSNet[8]	0.325	0.385	0.355	76.36	58.45	46.20	55.53	56.11	54.02	58.17	46.56	56.42
UCS-Net[9]	0.338	0.349	0.344	76.09	53.16	43.03	54.00	55.60	51.49	57.38	47.89	54.83
Vis-MVSNet[25]	0.369	0.361	0.365	77.40	60.23	47.07	63.44	62.21	57.28	60.54	52.07	60.03
CVP-MVSNet[10]	**0.296**	0.406	0.351	76.50	47.74	36.34	55.12	57.28	54.28	57.43	47.54	54.03
PatchmatchNet[12]	0.427	**0.277**	0.352	66.99	52.64	43.24	54.87	52.87	49.54	54.21	50.81	53.15
AA-RMVSNet[11]	0.376	0.339	0.357	77.77	59.53	51.53	**64.02**	**64.05**	59.47	60.85	**55.50**	61.51
EPP-MVSNet[26]	0.413	0.296	0.355	77.86	60.54	**52.96**	62.33	61.69	60.34	**62.34**	55.30	61.68
Ours	0.336	0.295	**0.316**	**80.39**	**65.64**	51.08	62.33	62.30	**61.89**	59.02	54.47	**62.14**

Table 2. **Ablation results on the proposed model.**

Method	2D ANNM	3D ANNM	R/D Refine	EA Loss	Acc.	Comp.	Overall.↓
Baseline (CasMVSNet [8])					0.325	0.385	0.355
+2D ANNM	✓				0.339	0.320	0.330
+2D&3D ANNM	✓	✓			0.343	0.306	0.325
+2D&3D ANNM+R/D Refine	✓	✓	✓		0.347	0.290	0.319
Proposed Method	✓	✓	✓	✓	0.336	0.295	0.316

that calculates the mean of precision and recall is reported. Tab. 1 shows the results of each method, where our model achieves the best or second-best in most scenes. The corresponding point cloud reconstructions are shown in Fig. 5.

3.4. Ablation Study

Tab. 2 shows the ablation experiments of our methods. The baseline model [8] uses fixed-size convolution for feature extraction, and incorrect cost matches are propagated through the coarse-to-fine flow. With the introduction of 2D ANNM, the non-local neighbors aware feature adaptively learns the spatial similarity matches and their correlations (Fig. 1(a)), resulting in a remarkable improvement in completeness. And the extended 3D ANNM further enhances the depth-wise perception ability. The use of pre-trained RGB guided depth refinement modules effectively corrects errors in coarser stages and prevents the accumulation of errors by introducing high-res RGB contour information (Fig. 1(b)). With the effect of a collaboration of all introduced modules, our proposed network obtains a 23.4% improvement in completeness while ensuring accuracy, achieving state-of-the-art results.

4. CONCLUSION

In this paper, we present the N2MVSNet which introduces the adaptive non-local neighbors matching strategy pixel-wise in the spatial domain and voxel-wise in the depth dimension. The proposed ANNM aggregates neighbors aware correlation by constructing non-local neighbors and corresponding weights, while the 3D ANNM further explores the depth perception among cost matches. And we adopt the pre-trained RGB guided depth refinement for depth hypothesis repolish in the coarsest stage, which prevents the accumulation and propagation of errors in the coarse-to-fine scheme. Finally, the energy aggregation loss is utilized for supervised training. Extensive results on the DTU and Tanks and Temples datasets demonstrate our state-of-the-art performance.

References

[1] Yasutaka Furukawa and Jean Ponce, "Accurate, dense, and robust multi-view stereopsis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007.

[2] Kiriakos N. Kutulakos and Steven M. Seitz, "A theory of shape by space carving," *International Journal of Computer Vision*, 1999.

[3] Johannes L. Schonberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys, "Pixelwise view selection for unstructured multi-view stereo," *european conference on computer vision*, 2016.

[4] Qingshan Xu, Weihang Kong, Wenbing Tao, and Marc Pollefeys, "Multi-scale geometric consistency guided and planar prior assisted multi-view stereo," 2022.

[5] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan, "Mvsnet: Depth inference for unstructured multi-view stereo," *european conference on computer vision*, 2018.

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," *medical image computing and computer assisted intervention*, 2015.

[7] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan, "Recurrent mvsnet for high-resolution multi-view stereo depth inference," *computer vision and pattern recognition*, 2019.

[8] Xiaodong Gu, Zhiwen Fan, Zuozhuo Dai, Siyu Zhu, Feitong Tan, and Ping Tan, "Cascade cost volume for high-resolution multi-view stereo and stereo matching," *computer vision and pattern recognition*, 2019.

[9] Shuo Cheng, Zexiang Xu, Shilin Zhu, Zhuwen Li, Li Erran Li, Ravi Ramamoorthi, and Hao Su, "Deep stereo using adaptive thin volume representation with uncertainty awareness," *arXiv: Computer Vision and Pattern Recognition*, 2019.

[10] Jiayu Yang, Wei Mao, Jose M. Alvarez, and Miaomiao Liu, "Cost volume pyramid based depth inference for multi-view stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[11] Zizhuang Wei, Qingtian Zhu, Chen Min, Yisong Chen, and Guoping Wang, "Aa-rmvsnet: Adaptive aggregation recurrent multi-view stereo network," *arXiv: Computer Vision and Pattern Recognition*, 2021.

[12] Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Pablo Speciale, and Marc Pollefeys, "Patchmatchnet: Learned multi-view patchmatch stereo," *computer vision and pattern recognition*, 2020.

[13] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl, "Large-scale data for multiple-view stereopsis," *International Journal of Computer Vision*, 2016.

[14] A. Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun, "Tanks and temples: benchmarking large-scale scene reconstruction," *ACM Transactions on Graphics*, 2017.

[15] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, "Feature pyramid networks for object detection," *arXiv: Computer Vision and Pattern Recognition*, 2016.

[16] Robert T. Collins, "A space-sweep approach to true multi-image matching," *computer vision and pattern recognition*, 1996.

[17] Jérémy Riviere, Paulo F. U. Gotardo, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler, "Single-shot high-quality facial geometry and skin appearance capture," *ACM Transactions on Graphics*, 2020.

[18] Wenzhe Shi, Jose Caballero, Ferenc Huszar, Johannes Totz, Andrew Peter Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," *computer vision and pattern recognition*, 2016.

[19] Zixiang Zhao, Jiangshe Zhang, Shuang Xu, Chunxia Zhang, and Junmin Liu, "Discrete cosine transform network for guided depth map super-resolution.," *arXiv: Computer Vision and Pattern Recognition*, 2021.

[20] Xiaofeng Wang, Zheng Zhu, Fangbo Qin, Yun Ye, Guan Huang, Xu Chi, Yijia He, and Xingang Wang, "Mvster: Epipolar transformer for efficient multi-view stereo," *arXiv preprint arXiv:2204.07346*, 2022.

[21] Rui Peng, Rongjie Wang, Zhenyu Wang, Yawen Lai, and Ronggang Wang, "Rethinking depth estimation for multi-view stereo: A unified representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8645–8654.

[22] B. Chitprasert and K. R. Rao, "Discrete cosine transform filtering," *Signal Processing*, 1990.

[23] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan, "Blendedmvs: A large-scale dataset for generalized multi-view stereo networks," *computer vision and pattern recognition*, 2019.

[24] Johannes L. Schonberger and Jan-Michael Frahm, "Structure-from-motion revisited," *computer vision and pattern recognition*, 2016.

[25] Jingyang Zhang, Yao Yao, Shiwei Li, Zixin Luo, and Tian Fang, "Visibility-aware multi-view stereo network.," *british machine vision conference*, 2020.

[26] Xinjun Ma, Yue Gong, Qirui Wang, Jingwei Huang, Lei Chen, and Fan Yu, "Epp-mvsnet: Epipolar-assembling based depth prediction for multi-view stereo," *international conference on computer vision*, 2021.