Cost Volume Pyramid Network with Multi-strategies Range Searching for Multi-view Stereo

Shiyu Gao[(✉)], Zhaoxin Li[(✉)] , and Zhaoqi Wang

Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
{gaoshiyu,zqwang}@ict.ac.cn, cszli@hotmail.com

Abstract. Multi-view stereo is an important research task in computer vision while still keeping challenging. In recent years, deep learning-based methods have shown superior performance on this task. Cost volume pyramid network-based methods which progressively refine depth map in coarse-to-fine manner, have yielded promising results while consuming less memory. However, these methods fail to take fully consideration of the characteristics of the cost volumes in each stage, leading to adopt similar range search strategies for each cost volume stage. In this work, we present a novel cost volume pyramid based network with different searching strategies for multi-view stereo. By choosing different depth range sampling strategies and applying adaptive unimodal filtering, we are able to obtain more accurate depth estimation in low resolution stages and iteratively upsample depth map to arbitrary resolution. We conducted extensive experiments on both DTU and BlendedMVS datasets, and results show that our method outperforms most state-of-the-art methods. Code is available at: https://github.com/SibylGao/MSCVP-MVSNet.git.

Keywords: Multi-view stereo · 3D reconstruction · Cost volume · Coarse-to-fine

1 Introduction

Multi-view stereo is one of the fundamental computer vision tasks which is widely used in augmented reality, 3D modeling and autonomous driving. In deep learning era, deep CNNs used for cost regularization and extracting representative image features have achieved promising performance. Yao et al. [2] first proposed an end-to-end MVS pipeline that constructs cost volume based on plane sweeping algorithm and aggregates different views by minimizing differential variance. However, this method consumes huge memory because that 3D CNN used for regularization is cubically proportional to image resolution. As a result, subsequent methods like [2,3] downsample high resolution images to regularize cost

volume in a smaller resolution. To this end, methods designed in coarse-to-fine manner [7–10] are put forward, which iteratively refine depth map based on cost volume pyramid and consume less memory.

However, current coarse-to-fine methods suffer from two limitations. First, the accuracy of the predicted depth map is highly dependent on the initial low-resolution depth map, since it is difficult to correct the depth of ill-posed and occluded pixels in the following narrow range. Second, current coarse-to-fine methods use same searching strategies in refinement stages after gaining initial depth map, which, however, not fully considered the characteristics of the cost volumes in each stage.
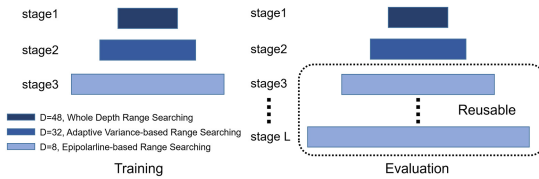

Fig. 1. Our method during training and evaluation.

In this work, we propose a multi-strategies cost volume pyramid multi-view stereo network (MSCVP-MVSNet). Instead of single depth range searching strategy, we utilize multi-dimensional information to calculate depth searching range for each layer. To further utilize the information contained in the cost volume, we introduce unimodal distribution as a training label at second stage during the training process.

Our main contributions can be summarized as follows:

We present multiple depth range searching methods in different stages of pyramid structure, leveraging multi-dimension information. On the second stage, variance-based strategy is applied to exploit previous predicted probabilities for each pixel. For the succeeding refinement stage, we employ parameter-free method to propagate neighboring information to an arbitrary resolution during upsampling.

To further exploit information in cost volume of deep MVS and obtain more accurate predictions in low-resolution stage before refinement, we propose unimodal assumption as a training label in second stage.

Quantitative results show that our method obtains SOTA results on DTU dataset and satisfactory qualitative results on BlendedMVS.

2   Related Work

2.1   Coarse-to-Fine MVS Methods

Deep MVS methods [3,4] based on pipeline of MVSNet [2] build cost volume at the resolution of output images, which usually occupy large memory dealing with high resolution dataset such as DTU [5] or Tanks and Temples [6]. In order to save memory and computation consumption, coarse-to-fine methods [8–10]are put forward. For example, CVP-MVSNet [9] and Cascade-MVSNet [10] build cost volume across the entire depth range in the coarsest resolution, after that, a narrowed sampling range is calculated based on previous depth predictions. Based on these works [9,10], Yu et al. [11] propose AACVP-MVSNet, which introduces attention mechanism to CVP-MVSNet [9] framework. Zhang et al. [12] took into account the visibility between different views based on Cascade-MVSNet [10].

2.2   Depth Sampling Range

Coarse-to-fine pyramid networks uniformly sample the entire depth range in the first stage. In the following stage, they iteratively narrow depth searching range by various strategies. CVP-MVSNet [9] determines the local sampling range around the current depth by back projecting the corresponding pixels along epipolar line in source views. Cas-MVSNet [10] narrows sampling range of each stage by hand-crafted range with specific decay ratio. For the first time, Cheng et al. [16] utilized variance of probability distribution to describe the uncertainty of depth estimation.

All these methods mentioned above employ identical sampling range searching strategies in each stage of three- or four-layer pyramid. In order to leverage both variance and neighbouring contextual information without adding complicated neural network modules, we apply different sampling range calculation strategies in different stage of coarse-to-fine MVS framework.

2.3   Cost Volume

Recently, cost volume is widely used in MVS and stereo matching methods. MVSNet [2] first introduces cost volume for end-to-end MVS pipeline by calculating photometric matching cost of each pixel in different fronto-parallel planes hypothesis. A standard cost volume has a resolution of $H \times W \times D \times F$, where H, W, D, F are height, width, number of plane hypothesis and feature channels, respectively. While cost volume indicates matching cost of each depth hypothesis of each pixel intuitively, it is regularized by 3D UNet to generate an estimated probability value and indirectly supervised as an intermediate layer. In order to integrate multi-scale information of cost volume, Shen et al. [18] proposed cost volume fusion module to obtain better initial disparity map. Like CFNet [18], we further utilize cost volume to obtain better initial depth map before refinement.
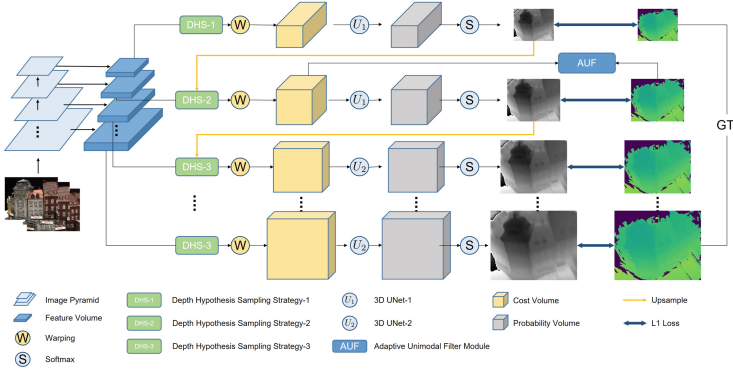
Fig. 2. The network structure of MSCVP-MVSNet.

3 Methods

3.1 Overview

In this section, we introduce our multi-strategies cost volume pyramid network for high-resolution MVS reconstruction in details. The overview of the network is shown in Fig. 2. We assume the input reference image denoted by $I_0 \in \mathbb{R}^{H \times W}$, and source images represented by $\{I_i\}_{i=1}^{N-1}$. To build a pyramidal structure, we downsample input images L times to obtain images pyramid $\{I_i^j\}_{j=1}^{L}$, where $i \in \{0, 1, \cdots, N\}$. Feature pyramid $\{F_i^j\}_{j=1}^{L}$ are build by weights-shared feature extraction module.

As shown in Fig. 2, three different sampling strategies and two separated UNets are employed in our framework.

Inspired by GwcNet [14], we build cost volume by group-wise correlation instead of calculating feature volume variance over all views proposed by Yao et al. [2].

3.2 Depth Sampling Range Estimation

As introduced in related work, previous methods [9,11,18] employ single strategy in each stage to calculate depth range, which either ignore statistical properties of each pixel or neighbouring information. To solve this, we fuse multi-dimensional information by simply combine different uncertainty estimation strategies in different stage without any additional neural network modules and achieve satisfactory results.

In this section, we present our depth hypothesis sampling strategies in details. As shown in Fig. 1, the number of pyramid layers in our framework is flexible, we train 3 different layers while evaluate with arbitrary number of stages.
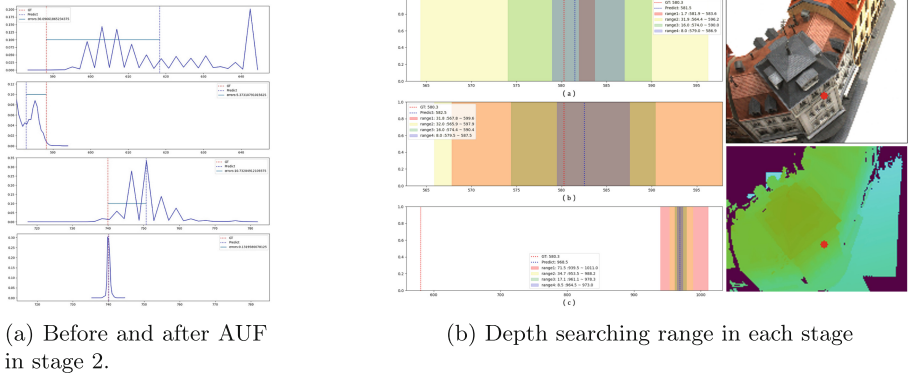
(a) Before and after AUF in stage 2.

(b) Depth searching range in each stage

Fig. 3. Depth searching range visualization.

In the first stage, we uniformly sampled depth hypothesis over the entire range to obtain a coarsest initial depth map. Due to the large sampling range, we sampled more depth hypothesis ($D^1 = 48$) in this stage.

For second stage, we take advantage of probability distributions to calculate specific depth sampling range of each pixel. Previous methods [13,18] indicate that texture-less and occluded pixels tend to have multiple or wrong matches, as a result, the expectation of the per-pixel distributions can not depict the properties of multimodality and dispersion. To solve this issue, we leverage the variance of the probability distribution as well as adaptive unimodal constraints (Sect. 3.3) to estimate per-pixel uncertainty and reduce local maxima of probabilities. We set the number of depth hypothesis, $D^2 = 32$ in this stage. For stage l, the variance at pixel i is defined as:

$$\hat{V}_i^l = \sum_{j=1}^{D^l} P_{i,j}^l (d_{i,j}^l - \hat{d}_i^l)^2, \tag{1}$$

where $P_{i,j}^l$ is the probability of pixel i at sampled depth j, $d_{i,j}^l$ is the depth of sampled plane j, $\hat{d}_i^l$ is the estimated depth of pixel i at current stage.

Different from UCSNet [16], we adopt the idea of CFNet [18] that originally proposed in stereo matching task, which use learned instead of hand-crafted scale parameters to determine confidence interval:

$$
\begin{aligned}
d_{max}^{l+1}(i) &= \hat{d}_i^l + \alpha^l \sqrt{\hat{V}_i^l} + \beta^l, \\
d_{min}^{l+1}(i) &= \hat{d}_i^l - \alpha^l \sqrt{\hat{V}_i^l} - \beta^l,
\end{aligned}
\tag{2}
$$

where α^l and β^l are learned parameters in stage l. Same as CFNet [18], we initial α^l and β^l as 0 at the beginning of training. In texture-less regions with multimodal distributions, the variances tend to be large, and adaptive uncertainty

range estimation algorithm adjust depth hypothesis to a larger range so as not to miss the truth depth value before small-range refinement. Depth searching ranges in Fig. 3 show the effectiveness of our variance-based method in 2nd stage.

Our first two layers have yielded fair results at the low resolution stage, and the depth values of high-resolution depth maps are obtained via upsampling operation. Specifically, we apply parameters-free method to determine sampling range, which takes advantage of contextual information provided by neighboring pixels along epipolar line [9].

3.3   Supervise on Cost Volume

As shown in Fig. 3, it is hard to correct misestimated depth in refinement stages. To attain better predictions before neighboring-based refinement, we further utilize the information in cost volume at 2nd stage.

Cost volume is defined to reflect the similarity between different views, where the true depth value should have the lowest cost, which means the probability distribution should be unimodal and peaked at the true depth hypothesis under ideal circumstances. Based on this assumption, we construct unimodal distributions as reference distributions which directly constraint on the cost volume to reduce errors introduced by multi-modal distributions. Following [13], we defined reference unimodal distribution as:

$$P^l(i) = softmax(-\frac{|d^l(i) - d_{gt}^l(i)|}{\sigma_i}), \tag{3}$$

where σ_i is variance of reference distribution for pixel i, which controls the sharpness of peak and it is defined as:

$$\sigma_i^l = \alpha_c^l(1 - f_i^l) + \beta_c^l, \tag{4}$$

where f_i^l is confidence value of pixel i in stage l. We estimate confidence value for each pixel by a 2D confidence estimation network. α_c^l and β_c^l are scale factor and lower bound, respectively. Different from [13], we use learned neural network parameters instead of hand-crafted factors to adapt different properties of probability distributions for different datasets. Large σ indicates low confidence of pixel, which usually caused by mismatch in textureless regions.

We leverage stereo focal loss proposed by AcfNet [13] to guide network to generate unimodal distributions for each pixel. The stereo focal loss is defined as:

$$\mathcal{L}_{SF} = \frac{1}{|\mathcal{P}|} \sum_{i \in \mathcal{P}} (\sum_{d=0}^{D-1} (1 - P_i(d))^{-\gamma} \cdot (-P_i(d) \cdot \log \hat{P}_i(d))), \tag{5}$$

where $P_i(d)$ is probability value of reference unimodal distribution at depth d of pixel i, and $\hat{P}_i(d)$ is estimated probability of pixel i at depth d given by our UNet. Instead of simple cross entropy loss, we set $\gamma \geq 0$ to force unimodal guidance to focus on high-confidence regions.

After adaptive unimodal filtering (AUF), some local maximas are eliminated, and the errors in stage 2 are decreased. Figure 3(left) presents depth searching range of our method with and without AUF, respectively. The depth sampling range in 2nd stage indicates that AUF narrows down the sampling range and contributes to a more accurate initial depth map before refinement.

3.4 Loss Function

Our total loss consists of three parts: regression loss in each stage, stereo focal loss and confidence loss, which is denoted as:

$$\mathcal{L} = \lambda_{SF}\mathcal{L}_{SF} + \lambda_C\mathcal{L}_C + \sum_{l=1}^{L} \omega^l \mathcal{L}_{regression}^l \tag{6}$$

where λ_{SF} and λ_C are two factors to balance stereo focal loss and confidence loss on second stage. The confidence loss $\mathcal{L}_C$ is defined as:

$$\mathcal{L}_C = \frac{1}{|\mathcal{P}|} \sum_{i \in \mathcal{P}} -\log f_i \tag{7}$$

We apply negative log-likelihood function as confidence loss to encourage confidence estimation network to predict high confidence values for each pixel.

Regression loss $\mathcal{L}_{regression}^l$ is defined to reflect the difference between the predicted depth map and ground-truth at stage l. We use hand-crafted weight ω at each stage. For stage l, the $L1$ norm is defined as:

$$\mathcal{L}_{regression}^l = \sum_{i \in \mathcal{P}} \|d_i^l - \hat{d}_i^l\|_1 \tag{8}$$

4 Experiment

4.1 Dataset

DTU Dataset. We train and evaluate our network on DTU dataset [5] to obtain quantitative results. DTU dataset [5] consists of 124 large scale of scenes in 49 or 64 different views and 7 different light conditions, with the evaluation reference obtained by a structured light scanner. We use the same splited training and evaluation sets with [3,9,11]. While the original size of evaluation image is 1600×1200, we crop it to 1600×1184 to fit the upsample process.

BlendedMVS. BlendedMVS [22] is a collection of images captured from different views of 113 various scenarios. It contains 17K training samples in low-resolution (768×576) as well as high-resolution (2048×1536). Following the official training and validation list given by the released dataset files, we divided 106 scenes for training and the other 7 for validation in low-resolution Blended-MVS. We train our model on low-resolution BlendedMVS and evaluate on both low-resolution and high-resolution.

Table 1. Quantitative results on DTU

Methods	acc.	comp.	overall
MVSNet [2]	0.396	0.527	0.462
R-MVSNet [3]	0.383	0.452	0.418
MVSCRF [4]	0.371	0.426	0.398
PointMVSNet [7]	0.361	0.421	0.391
CVP-MVSNet [9]	**0.296**	0.406	0.351
AACVP-MVSNet [11]	0.357	0.326	0.341
Vis-MVSNet [12]	0.369	0.361	0.365
USCNet [16]	0.338	0.349	0.344
PVSNet [21]	0.337	0.315	**0.326**
Ours	0.379	**0.278**	0.328

Table 2. Different strategies

Strategies	acc.	comp.	overall
DHS1	0.444	0.361	0.402
DHS1+DHS2	**0.338**	0.349	0.344
DHS1+DHS3	0.404	0.321	0.362
DHS1+DHS2+DHS3	0.389	**0.279**	**0.334**

Note: DHS1 denotes uniformly sampling the whole range, DHS2 denotes variance-based method, and DHS3 back-projects pixels along epipolar line to calculate depth searching range.

4.2   Implementation Details

Training. We train and evaluate our model on DTU dataset and low-resolution BlendedMVS. For first stage, we uniformly sample the whole depth range $[425, 1065]$ with $D^1 = 48$, while for 2nd and 3rd stage, we choose $D^2 = 32$ and $D^3 = 8$, respectively. As the training process with high-resolution inputs is memory and time consuming, we downsample the training set into a size of 320×256, and the coarsest resolution is 40×32 in the first stage. We set hyper-parameters $\lambda_{SF} = 10$, $\lambda_C = 80$ in equation (6) and choose $\omega^1 = 0.5$, $\omega^2 = 1$, $\omega^3 = 2$ to balance $L1$ loss in each stage. As for the reference unimodal distribution, the scale factors are initialized as $\alpha_c^2 = 13$ and $\beta_c^2 = 9$, respectively, based on empirical evidence from [13]. We use 3 different views as inputs and Adam [19] as optimizer in the training stage of the proposed network. We set batch size as 16 and train our model on 2 Nvidia GeForce RTX 3090 for 40 epoches with initial learning rate 0.001 multiplied by 0.5 at 10th, 12th, 14th, 20th epoch.

Evaluation. For DTU dataset, we crop the original images to 1600×1184 for evaluation. We set $L = 5$ for image feature pyramid to maintain a similar size with training stage at the coarsest stage (50×37). Similar to [2,3,9], we choose 5 views in evaluation for fair comparison. The depth sampling numbers D in each stage the same as training process. As for BlendedMVS, we evaluate our proposed method on both low-resolution and high-resolution dataset.

Post processing and Metrics. After estimating the depth map, we fuse all views into a dense point cloud model for each scene. For fair comparison, we follow the common post processing method used by [2,3,9], which is a fusion method provided by Galliani et al. [20]. We run the official evaluation code provided by DTU dataset [5] to obtain quantitative results of mean accuracy (acc.), mean completeness (com.) and overall score (overall). The evaluation results are listed in Table 1.

Fig. 4. Reconstruction results on DTU dataset.

4.3 Results on DTU Dataset

We train and evaluate our method on DTU dataset to conduct quantitative results in comparison with other learning based methods. As shown in Table 1, our method achieves state-of-the-art results in overall score, which is comparable to PVSNet [21]. Especially, our method outperforms all methods in Table 1 in terms of completeness. As shown in Fig. 4, We visualize several reconstructed 3D models constructed by CVP-MVSNet [9], AACVP-MVSNet [11] and our proposed method.

4.4 Results on BlendedMVS

As BlendedMVS dataset does not provide any reference point clouds for quantitative evaluation, we conduct the visual comparison with CVP-MVSNet [9]. $L = 3$ in training process, while for evaluation, we set $L = 5$ and $L = 6$ for low and high resolution evaluation sets, respectively. In the same way, we compare our method with CVP-MVSNet [9] and the results of low- and high-resolution dataset are shown in Fig. 5. On high-resolution data sets, the superiority of our method in terms of completeness is even more evident.

4.5 Effectiveness of Multi-strategies

As shown in Table 2, we compare our proposed multi-strategies with other combinations of strategies. For strategy "DHS1", we apply DHS1 and uniformly sample at each stage with handcrafted searching range $[40, 20, 10, 5]$ from 2nd to 5th stage (range 1 corresponds to stage 2). "DHS1+DHS2" and "DHS1+DHS3" perform results of single strategy DHS2 and DHS3 from 2nd to 5th stage, respectively.

As shown in Fig. 3, CVPMVSNet [9] applies DHS3 in each stage and fails to locate an interval which contains true depth value from 2nd to the last stage. Its single and inflexible range searching strategy makes it hard to jump out of the pattern and rectify mismatch in previous stage. We believe that DHS2 which is based on the variance of previous prediction is more accurate and effective to locate true depth value (see (b) range 1 in Fig. 3), but proper scale factors are needed in each specific stage. Our proposed multi-strategies method combines both DHS2 and DHS3, in the second stage, DHS2 gives a reasonable searching range based on previous predicted probabilities, while for the rest stages, DHS3

which is parameter-free provides an effective way to propagate depth of neighboring pixels along epipolar line to an arbitrary resolution during upsampling refinement. Noteworthy, multi-strategies method works better when it combines with two separated UNets (see CVP-MS and CVP-MS-U^2Net in Table 2).

Fig. 5. Reconstruction results on BlendedMVS dataset.

4.6   Ablation Study

In this section, we perform ablation experiments on DTU dataset to validate the effectiveness of each component of our proposed network. Results are shown in Table 3. Below we analyse each component in details.

Table 3. Ablation study on DTU dataset

Methods	Variance	Epipolar line	U^2Net	Auf	acc.	comp.	overall
CVP (baseline)	✗	✓	✗	✗	**0.313**	0.394	0.354
CVP-MS	✓	✓	✗	✗	0.343	0.439	0.391
CVP-U^2Net	✗	✓	✓	✗	0.330	0.379	0.355
CVP-MS-Auf	✓	✓	✗	✓	0.321	0.398	0.360
CVP-MS-U^2Net	✓	✓	✓	✗	0.389	0.279	0.334
Ours	✓	✓	✓	✓	0.379	**0.278**	**0.328**

- **Non-parameter-sharing UNet.** 3D UNet is designed for cost volume regularisation and explore cost volume information in three dimensions. Quantitative results on DTU dataset show that our two parameter-separating UNets gain better results (0.328 vs.0.360) than parameter-sharing UNet. The huge gap indicates that former stages which search in a wider range have different characteristics with refinement stages in the cost volume regularization process.
- **Supervise on cost volume.** While multi-strategies with two non-parameter-sharing UNet framework has achieve promising results (see CVP-MS-U^2Net in Table 3), we obtain even better results when further adding adaptive unimodal filtering (AUF) on 2nd stage. As shown in Fig. 3, the depth sampling range in 2nd stage is narrowed after adding AUF module. Interestingly, quantitative results of CVP-MS-Auf and CVP-MS in Table 3 show that

adaptive unimodal filtering gives a greater boost when parameter-sharing UNet is adopted.

- **Image resolution during training and evaluation.** Table 4 shows that the performance of the model trained with higher resolution input is better than that with lower resolution input. To discover the relationship between pyramid levels and quality of output depth map, we also evaluate our method with different pyramid levels on DTU dataset. As shown in Table 4, coarse-to-fine network with 5 pyramid stages achieves the best overall score.

Table 4. Quantitative results on DTU dataset with different training and evaluation resolution.

Coarsest Res_T	Coarsest Res_E	$Levels_E$	mem. (M)	runtime (s)	acc.	comp.	overall
40×32	25×18	6	**6809**	2.543	0.372	0.292	0.332
20×16	25×18	6			0.382	0.324	0.353
40×32	50×37	5	7863	2.550	0.379	**0.278**	**0.328**
20×16	50×37	5			0.371	0.328	0.349
40×32	100×74	4	6935	2.483	0.360	0.311	0.335
20×16	100×74	4			**0.349**	0.478	0.413
40×32	200×148	3	7861	**2.366**	0.375	0.530	0.452
20×16	200×148	3			0.531	1.959	1.245

5   Conclusion

In this paper, we present an efficient deep-learning based cost volume pyramid network for MVS. By combining different sampling range estimation strategies for each stage, we integrate multi-dimensional information without additional neural network modules. Then, we apply adaptive unimodal filters to further improve the low-resolution depth map before refinement. Results on different datasets show the effectiveness and generalisability of our method.

References

1. Ji, M., Gall, J., Zheng, H., Liu, Y., Fang, L.: 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)
2. Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L.: MVSNet: depth inference for unstructured multi-view stereo. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11212, pp. 785–801. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01237-3_47

3. Yao, Y., Luo, Z., Li, S., Shen, T., Fang, T., Quan, L.: 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)
4. Xue, Y., et al.: MVSCRF: learning multi-view stereo with conditional random fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4312–4321 (2019)
5. Aanas, H., Jensen, R.R., Vogiatzis, G., Tola, E., Dahl, A.B.: Large-scale data for multiple-view stereopsis. Int. J. Comput. Vis. **120**(2), 153–168 (2016)
6. Knapitsch, A., Park, J., Zhou, Q.Y., Koltun, V.: Tanks and temples: benchmarking large-scale scene reconstruction. ACM Trans. Graph. (ToG) **36**(4), 1–13 (2017)
7. Chen, R., Han, S., Xu, J., Su, H.: Point-based multi-view stereo network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1538–1547 (2019)
8. Yu, Z., Gao, S.: Fast-MVSNet: sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1949–1958 (2020)
9. Yang, J., Mao, W., Alvarez, J.M., Liu, M.: 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)
10. Gu, X., et al.: 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)
11. Yu, A., et al.: Attention aware cost volume pyramid based multi-view stereo network for 3D reconstruction. ISPRS J. Photogramm. Remote Sens. **175**, 448–460 (2021)
12. Zhang, J., Yao, Y., Li, S., Luo, Z., Fang, T.: Visibility-aware multi-view stereo network. arXiv:2008.07928, August 2020. https://arxiv.org/abs/2008.07928
13. Zhang, Y., et al.: Adaptive unimodal cost volume filtering for deep stereo matching. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 07, pp. 12926–12934, April 2020
14. Guo, X., Yang, K., Yang, W., Wang, X., Li, H.: Group-wise correlation stereo network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3273–3282 (2019)
15. Chang, J. R., Chen, Y.S.: Pyramid stereo matching network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5410–5418 (2018)
16. Cheng, S., 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)
17. Mao, Y., et al.: UASNet: uncertainty adaptive sampling network for deep stereo matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6311–6319 (2021)
18. Shen, Z., Dai, Y., Rao, Z.: CFNet: cascade and fused cost volume for robust stereo matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13906–13915 (2021)
19. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
20. Galliani, S., Lasinger, K., Schindler, K.: Massively parallel multiview stereopsis by surface normal diffusion. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 873–881 (2015)

21. Xu, Q., Tao, W.: PVSNet: pixelwise visibility-aware multi-view stereo network. arXiv preprint arXiv:2007.07714 (2020)
22. Yao, Y., 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)