Multi-View Stereo Representation Revist: Region-Aware MVSNet

Yisu Zhang[1],   Jianke Zhu[1,2*] Lixiang Lin[1],
[1]Zhejiang University
[2]Alibaba-Zhejiang University Joint Research Institute of Frontier Technologies
{zhyisu, jkzhu, lxlin}@zju.edu.cn

Abstract

Deep learning-based multi-view stereo has emerged as a powerful paradigm for reconstructing the complete geometrically-detailed objects from multi-views. Most of the existing approaches only estimate the pixel-wise depth value by minimizing the gap between the predicted point and the intersection of ray and surface, which usually ignore the surface topology. It is essential to the textureless regions and surface boundary that cannot be properly reconstructed. To address this issue, we suggest to take advantage of point-to-surface distance so that the model is able to perceive a wider range of surfaces. To this end, we predict the distance volume from cost volume to estimate the signed distance of points around the surface. Our proposed RA-MVSNet is patch-aware, since the perception range is enhanced by associating hypothetical planes with a patch of surface. Therefore, it could increase the completion of textureless regions and reduce the outliers at the boundary. Moreover, the mesh topologies with fine details can be generated by the introduced distance volume. Comparing to the conventional deep learning-based multi-view stereo methods, our proposed RA-MVSNet approach obtains more complete reconstruction results by taking advantage of signed distance supervision. The experiments on both the DTU and Tanks & Temples datasets demonstrate that our proposed approach achieves the state-of-the-art results.

1. Introduction

Multi-view stereo (MVS) is able to efficiently recover geometry from multiple images, which makes use of the matching relationship and stereo correspondences of overlapping images.

To achieve the promising reconstruction results, the conventional patch-based and PatchMatch-based methods [2, 11, 27] require rich textures and restricted lighting condi-
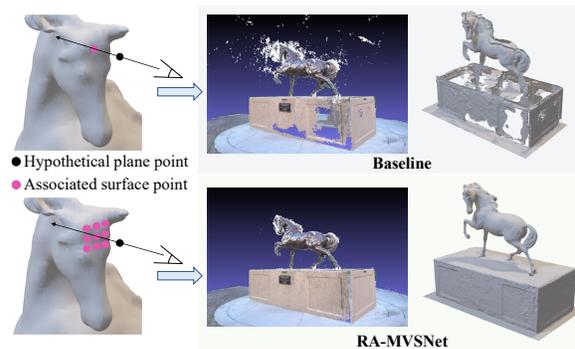

Figure 1. **Comparison on reconstruction results between baseline and RA-MVSNet.** Our RA-MVSNet enables the model to perceive a wider range of surfaces so as to achieve the promising performance in complementing textureless regions and removing outliers at boundaries. Furthermore, our model is able to generate correct mesh topologies with fine details based on estimated point-to-surface distances of spatially sampled points.

tions. Alternatively, the deep learning-based approaches [4, 14, 15, 40] try to take advantage of global scene semantic information, including environmental illumination and object materials, to maintain high performance in complex lighting. The key of these methods is to warp deep image features into the reference camera frustum so that the 3D cost volume can be built via differentiable homographies. Then, the depth map is predicted by regularizing cost volume with 3D CNNs.

Despite the encouraging results, the pixel-wise depth estimation suffers from two intractable flaws. One is the low estimation confidence in the textureless area. The other is many outliers near the boundary of the object. This is mainly because the surface is usually treated as a set of uncorrelated sample points rather than the one with topology. As each ray is only associated with a single surface sampling point, it is impossible to pay attention to the adjacent area of the surface. As shown in Fig. 1, the estimation of each depth value is constrained by only one surface sample point, which makes it unable to use the surrounding surface for inference. Unfortunately, it is difficult to infer without

*Corresponding author is Jianke Zhu.

broader surface information in textureless regions and object boundaries. Therefore, too small perception range limits the existing learning-based MVS methods.

To tackle this issue, we present a novel RA-MVSNet framework that is able to make each hypothetical plane associated with a wider surface area through point-to-surface distance. Thus, our presented method is capable of inferring the surrounding surface information at textureless areas and object boundaries. To this end, our network not only estimates the probability volume but also predicts the point-to-surface distance of each hypothetical plane. Specifically, RA-MVSNet makes use of the cost volume to generate the probability and distance volumes, which are further combined to estimate the final depth map. The introduction of point-to-surface distance supervision uses the model patch-aware in estimating the depth value corresponding to a particular pixel. This leads to the improved performance in textureless or boundary areas. Since the distance volume estimates the length of the sample points near the surface, we are able to predict a SDF-based implicit representation with the correct topology and fine details.

In summary, our contribution is three-fold:

- We introduce point-to-surface distance supervision of sampled points to expand the perception range predicted by the model, which achieves complete estimation in textureless areas and reduce outliers in object boundary regions.

- To tackle the challenge of lacking the ground-truth mesh, we compute the signed distance between point sets based on the triangulated mesh, which trades off between accuracy and speed.

- Experimental results on the challenging MVS datasets show that our proposed approach performs the best both on indoor dataset *DTU* [1] and large-scale outdoor dataset *Tanks and Temples* [17].

2. Related Work

We review the multi-view stereo studies from two aspects, including conventional methods and learning-based approaches.

Conventional Multi-View Stereo. The conventional MVS methods make use of various 3D representations, such as mesh [9], point cloud [11, 19], voxel [18, 28] and depth map [3, 12, 27]. Among these different representations, the depth map-based methods can obtain more complete surface reconstruction with higher robustness. They avoided solving the intractable topology problem by formulating the multi-view reconstruction into a depth estimation problem, which fuses all depth map into single 3D point cloud. Among them, COLMAP [27] and ACMM [35] can obtain the stable results. Specifically, ACMM employs multi-scale

geometric consistency to reconstruct features at different scales. COLMAP estimates the pixelwise depth and normal using photometric and geometric priors. In the cases of complicated scenario, large matching noise and poor correspondences, the results of traditional MVS may have the obvious artifacts.

Learning-based Multi-View Stereo. To address the limitations of traditional methods, deep learning-based approaches are proposed to robustly estimate depth map. MVSNet [40] firstly builds 3D cost volume to aggregate the warped features from the reference and source images, and then regresses the depth map by a 3D CNN. Later, some works [5, 8, 45] take into consideration of the attention mechanism to focus on the areas to be reconstructed. [23, 30] replace the cost volume representation to reduce the unnecessary computation. Ding et al. [7] try to optimize the depth map representation. Meanwhile, some studies intend to learn how to regularize cost volume better by hybrid 3D U-Net [20, 29] and epipolar attention [22, 39]. Although the vanilla MVSNet is able to obtain the pixel-wise depth prediction, dense hypothetical planes and 3D cost volume consume a large amount of memory. Recurrent MVSNet architectures [33, 34, 41], coarse-to-fine manner [13] and multi-stage binary search [23] are proposed to further excavate the potential capacity of this pipeline. Similar ideas are later explored to reduce the memory consumption of 3D convolutions and increase the depth quality, such as coarse-to-fine depth optimization [6, 36, 37, 44], attention-based feature aggregation [8, 21, 32, 43, 46], and patch matching-based method [31]. Meanwhile, Uni-MVSNet [25] analyzes the impact of regression and classification operations in the pipeline, and combines these two approaches to achieve more accurate predictions.

In general, the traditional methods based on patch matching cannot cope with the complicated lighting conditions and textureless areas while the deep learning methods based on cost volume predict many outliers in the object's boundary regions. In this paper, our proposed RA-MVSNet makes learning-based MVS patch-aware to increase the prediction performance for textureless regions and reduce erroneous outlier points near the object's boundary.

3. Method

In this section, we introduce the detailed structure of the proposed RA-MVSNet. As shown in Figure 2, the overall framework mainly consists of three parts, including cost volume construction, the multi-scale depth map and signed distance prediction, and handling of ground truth. Since our proposed point-to-surface distance supervision employs an additional branch to calculate the signed distance of the sampling points around the surface through cost volume, it is easy to be added into the existing learning-based MVSNet scheme with slight changes. To investigate the effective-
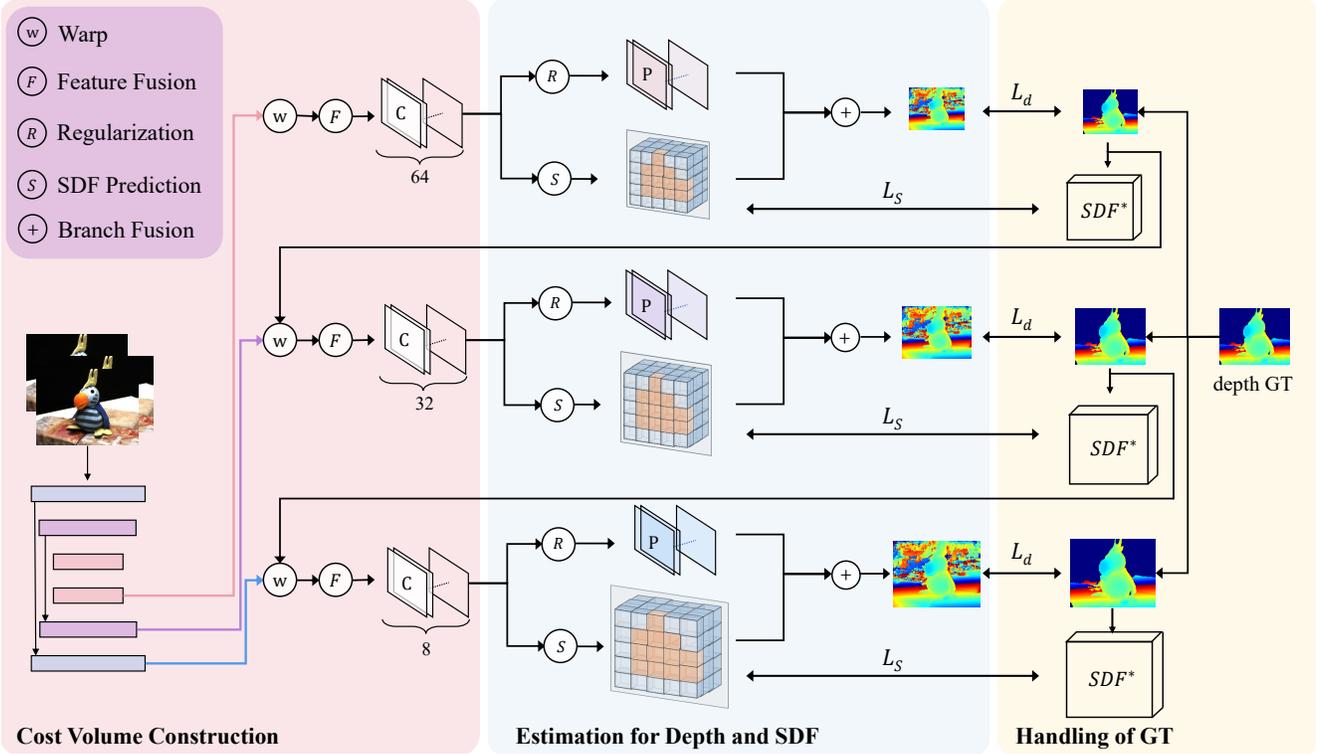
Figure 2. **Illustration of RA-MVSNet.** Our RA-MVSNet framework consists of two branches. The first branch predicts probability volume, and the second one estimates the signed distance volume. Fusing two branches can get the filtered depth maps while SDF branch can generate implicit representation.

ness of our method, we mainly employ the cascade MVS-Net as the baseline and use two branches on the basis of Cas-MVSNet [13] to predict the depth and signed distance, respectively.

3.1. Cost Volume Construction

The construction of cost volume mainly relies on MVS-Net framework [40], which utilizes the warped frustum features to predict the depth map $\{\mathbf{D} \in \mathbb{R}^{H' \times W'}\}$ corresponding to the reference image $\{\mathbf{I_0} \in \mathbb{R}^{H' \times W'}\}$. The feature volume is aggregated by warping source image features into the reference view, where all image features are extracted by 2D FPN-based network with the shared weights. To pay more attention to the object to be reconstructed, we employ Recursive Feature Pyramid (RFP) structure [26] as image encoder in order to obtain a pyramid of feature maps $\{\mathbf{F}_i \in \mathbb{R}^{C' \times H' \times W'}\}_{i=1}^{N}$ with three different scales.

By regularizing 3D cost volume in the whole known depth range, the estimated depth map consists of the depth hypothesis of M layers. The key to learning-based MVS lies in the way of dealing with 3D cost volume. Specifically, the feature volume can be aggregated by differentiable homography as below

$$\mathbf{H}_i(d) = d\mathbf{K}_i\mathbf{R}_i\mathbf{R}_0^{-1}\mathbf{K}_0^{-1}, \qquad (1)$$

where d refers to the hypothetical depth of source image

$\mathbf{R}_i$. $\mathbf{K}_i\mathbf{R}_i$ and $\mathbf{K}_0\mathbf{T}_0$ represent camera intrinsic and extrinsic parameters of source and reference images, respectively. Therefore, the warped pixel p' in source image $\mathbf{I}_i$ of reference pixel p can be computed by

$$\mathbf{p}' = \mathbf{H}_i(d) \cdot p + \mathbf{t}_{\mathbf{I}_0 \to \mathbf{I}_i}, \qquad (2)$$

where $\mathbf{t}_{\mathbf{I}_0 \to \mathbf{I}_i}$ is the relative camera translation from reference to the source image. To handle the arbitrary number of source images, we aggregate all feature volumes $\{V_i \in \mathbb{R}^{D \times C' \times H' \times W'}\}_{i=0}^{N-1}$ into single cost volume $\{\mathbf{C} \in \mathbb{R}^{D \times C' \times H' \times W'}\}$ using an adaptive strategy so that several 3D CNN layers can be employed to predict the pixel-wise weighting matrices $\{W_i\}_{i=1}^{N-1}$. Thus, the final cost volume can be computed as follows

$$\mathbf{C} = \sum_{i=1}^{N-1} \frac{1}{N-1} \mathcal{W}_i \odot (\mathbf{V}_i - \mathbf{V}_0)^2, \qquad (3)$$

where $\mathbf{C}$ is the cost volume of reference view. $\odot$ denotes the element-wise multiplication. $\mathbf{V}_i$ and $\mathbf{V}_0$ are the features extracted from source images and reference view using image encoder.

3.2. Signed Distance Supervision

The point-to-surface distance is usually represented as SDF (signed distance field) in a recent study [24]. The

core of this implicit representation is to calculate the distance from the sampled point near the surface to the object. Therefore, we follow the idea of SDF to construct a distance volume to predict the point-to-surface distances so as to take advantage of implicit representation.
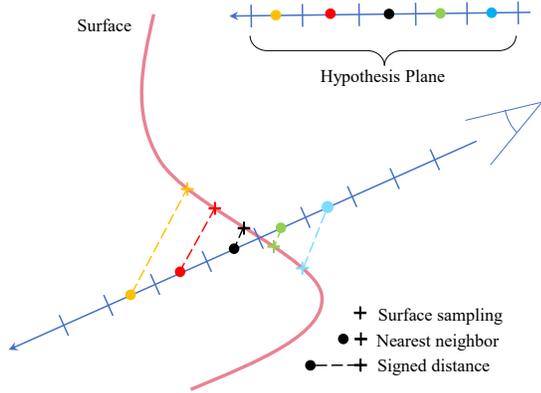

Figure 3. **Ground truth of signed distance.** The ground truth of signed distance is represented by the sided distance between the two point sets. We treat each hypothetical plane as a sampled point around the surface and find its corresponding nearest-neighbor surface sampled point to get the ground truth signed distance.

Given the 3D cost volume aggregating the feature of reference view and source views, the regularization networks are typically employed to obtain the probability volume $\mathbf{P}$ that is treated as the weight of hypothetical planes at different depths

$$\mathbf{P} = F_{softmax}(\mathbf{C}), \quad (4)$$

where $F_{SoftMax}$ is softmax-based 3D CNN regularization networks. Distance volume $\mathbf{S}$ represents the signed distance of these hypothetical planes

$$\mathbf{S} = F_{tanh}(\mathbf{C}), \quad (5)$$

where F_{tanh} denotes the tanh-based 3D CNN regularization networks. As the points far away from the surface are usually unhelpful for reconstruction, we employ $tanh$ as the activation layer for distance volume. Thus, we can focus on nearby sampled points.

Since the predictions of distance are introduced, we need to extend the ground truth from depth map to the signed distance field. Therefore, the depth map only contains the sampled points with the signed distance of 0, which lacks the ground truth of points around the surface. For an exact query point $\mathbf{p}_i$ from each hypothetical plane of cost volume $\{\mathbf{C}\}$, we compute the shortest distance from $\mathbf{p}_i$ to the surface sampling points $\mathbf{p}'$ as the ground truth of the signed distance. As shown in Fig. 3, we employ the two-point distance $d(\mathbf{p}_i, \mathbf{p}'_j)$ that is calculated by Kaolin [10] as

ground truth signed distance. To speed up the process, finding the nearest neighbor from all surface sampling points is replaced by local search, as shown in Fig. 4.
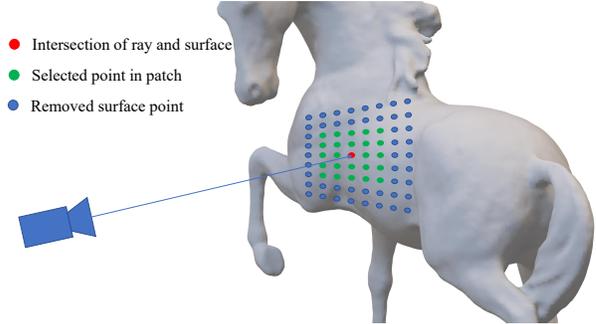

Figure 4. **Patch-based nearest neighbor search.** The nearest neighbor is usually located near the query point so that a large number of useless surface sampled points are removed while only the sampled points within the local patch located at the intersection are retained.

This patch-based local search method keeps the points that need to be calculated as few as possible within a reasonable range, thereby reducing the time complexity on search. We assume that the resolution of the depth map is $H \times W$ and the number of query points is n. Then, the time complexity of naive calculation is $O(n \times H \times W)$, which is proportional to the resolution of the depth map. In contrast, the time complexity of patch-based local search is simplified to $O(n \times k \times k)$, where k is the patch size and usually set to 5. Therefore, the time complexity of patch-based local search can be simplified to $O(n)$. That is to say, it is only proportional to the number of query points n, and the search time for each query point is constant.

3.3. Volume Fusion

Once the probability volume $\{\mathbf{P} \in \mathbb{R}^{D \times H' \times W'}\}$ and distance volume $\{\mathbf{S} \in \mathbb{R}^{D \times H' \times W'}\}$ are obtained, we fuse these two volumes to get the final depth map $D \in \mathbb{R}^{H' \times W'}$. In general, a softmax-based regularization network is typically employed to predict the depth map from $\mathbf{P}$ that is treated as the weight of hypothetical planes at different depths. Therefore, the depth map can be calculated as follows

$$D^{U,V} = \sum_{i=\mathbf{d}_{min}^{U,V}}^{\mathbf{d}_{max}^{U,V}} i\mathbf{P}(i)^{U,V}, \quad (6)$$

where $\mathbf{d}_{min}$ and $\mathbf{d}_{max}$ refer to the distance of the nearest and farthest hypothetical plane, respectively. However, this method has the accuracy problems due to involving with multiple invalid planes in the calculation. The depth value of a pixel (U, V) is only related to several hypothetical planes corresponding to this pixel, which cannot be associated with other sampled points on surface.
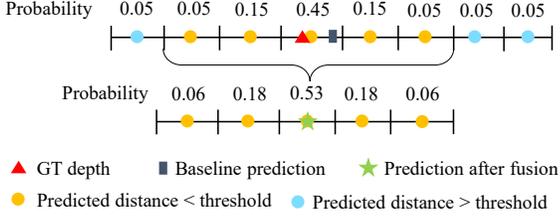
Figure 5. **Illustration of volume fusion.** RA-MVSNet can remove the invalid hypothesis planes through distance volume, which makes the results more accurate.

Algorithm 1: Branches Fusion

Input: Probability volume $\mathbf{P} \in \mathbb{R}^{D' \times H' \times W'}$;
         Distance volume $\mathbf{S} \in \mathbb{R}^{D' \times H' \times W'}$.
Output: Depth map $\mathbf{D} \in \mathbb{R}^{H' \times W'}$.
Initialization: Depth map $\mathbf{D} = 0$.

1 **for** $(u, v) = (1, 1)$ **to** (H', W') **do**
2   **for** $i = \mathbf{d}_{min}^{u,v}$ **to** $\mathbf{d}_{max}^{u,v}$ **do**
3   **if** $\mathbf{S}_i^{u,v} \leq threshold$ **then**
4           $\mathbf{D}_i^{u,v} = \mathbf{D}_i^{u,v} + Softmax(\mathbf{P}_i^{u,v}) \times \mathbf{d}_i^{u,v}$
5   **else**
6           $\mathbf{D}_i^{u,v} = \mathbf{D}_i^{u,v}$;
7   **end**
8   **end**
9 **end**
10 **return** $\mathbf{D}$.

As shown in Fig. 5, we fuse probability volume $\mathbf{P}$ and the introduced distance volume $\mathbf{S}$ to calculate the depth map so that each pixel is related to the surrounding surface patch. Specifically, $\mathbf{S}$ can be regarded as a filter of probability values by a threshold. The fusion process of these two volumes is illustrated in Algorithm 1. Finally, we use depth map ground truth and generated signed distance ground truth for supervision of two volumes P and S. We employ L_1 loss for depth map and signed distance as follows

$$L_d = \sum_{i=1}^{3} \|D_i^* - D_i\|, \tag{7}$$

$$L_S = \sum_{i=1}^{3} \|S_i^* - S_i\|, \tag{8}$$

where D_i^* and S_i^* are ground truth depth map and point-to-surface distance at stage i, respectively. D_i and S_i are the predicted value for two branches. Therefore, the total loss L of our model is the weighted sum of two branches:

$$L = L_d + \lambda \cdot L_S, \tag{9}$$

λ is a weight to balance two terms, which is set to 0.1 in all experiments.

3.4. Supervision of SDF Branch

Since we generate the point-to-surface distance ground truth from corresponding depth map, the error bound analysis is necessary. A reasonable assumption is to employ the triangulated mesh to represent the surface. There are three difference cases, as shown in Fig. 6.
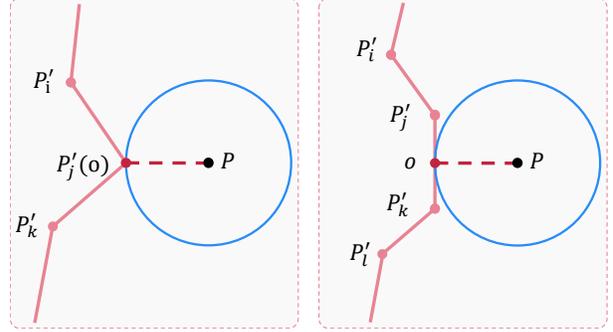
(a) The tangent point of the ball coincides with the surface sampled point

(b) The tangent point of the ball falls on one of the sides of the triangle
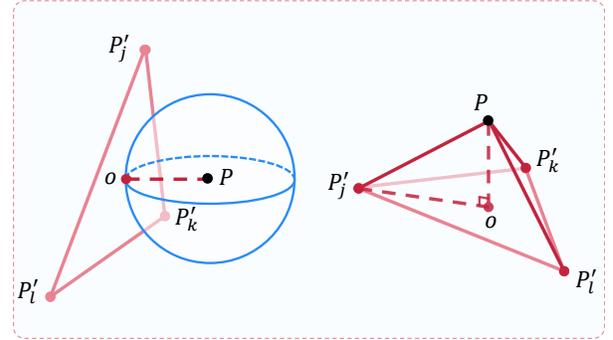
(c) The tangent point of the ball is inside the triangle

Figure 6. **Three cases for approximation.** The basic assumption is that the surface is represented by a triangular patch, which can only appear either straight or flat, rather curved.

In case (a), the largest sphere centering at the query point $\mathbf{p}$ is tangent to the surface of the object at point $\mathbf{o}$. Then, the ground truth of signed distance at $\mathbf{p}$ is $d(\mathbf{p}, \mathbf{o})$. The sided distance from the query point $\mathbf{p}$ to the sampled point set $\{\mathbf{p}'\}$ is $d(\mathbf{p}, \mathbf{p}_j')$. Since $\mathbf{p}_j'$ coincides with the tangent point $\mathbf{o}$, the error for case (a) is e_a^2 as below

$$\mathbf{e}_a^2 = \left(d(\mathbf{p}, \mathbf{o}) - d\left(\mathbf{p}, \mathbf{p}_j'\right)\right)^2 = 0, \tag{10}$$

where $d(\mathbf{p}, \mathbf{o})$ and $d\left(\mathbf{p}, \mathbf{p}_j'\right)$ represent ground truth and the approximate value of signed distance, respectively.

In case (b) and (c), we use the similar analytical method. Suppose $\mathbf{o}'$ and $\mathbf{o}''$ are the tangent points of the surface and the sphere centering at $\mathbf{p}$. The ground truth of signed distance in case (b) is $d(\mathbf{p}, \mathbf{o}')$, which is $d(\mathbf{p}, \mathbf{o}'')$ in case (c).

Method	Intermediate									Advanced						
	Mean	Fam.	Fra.	Hor.	Lig.	M60	Pan.	Pla.	Tra.	Mean	Aud.	Bal.	Cou.	Mus.	Pal.	Tem.
MVSNet [40]	43.48	55.99	28.55	25.07	50.79	53.96	50.86	47.90	34.69							
Point-MVSNet	48.27	61.79	41.15	34.20	50.79	51.97	50.85	52.38	43.06	-	-	-	-	-	-	-
CVP-MVSNet [38]	54.03	76.50	47.74	36.34	55.12	57.28	54.28	57.43	47.54	-	-	-	-	-	-	-
P-MVSNet [5]	55.62	70.04	44.64	40.22	**65.20**	55.08	55.17	60.37	54.29	-	-	-	-	-	-	-
D^2HC-RMVSNet [37]	59.20	74.69	56.04	49.42	60.08	59.81	59.61	60.04	53.92	-	-	-	-	-	-	-
RayMVSNet [33]	59.48	78.55	61.93	45.48	57.59	61.00	59.78	59.19	52.32	-	-	-	-	-	-	-
PatchmatchNet [31]	53.15	66.99	52.64	43.24	54.87	52.87	49.54	54.21	50.81	32.31	23.69	37.73	30.04	41.80	28.31	32.29
CasMVSNet [13]	56.84	76.37	58.45	46.26	55.81	56.11	54.06	58.18	49.51	31.12	19.81	38.46	29.10	43.87	27.36	28.11
AA-RMVSNet [32]	61.51	77.77	59.53	51.53	64.02	64.05	59.47	60.85	55.50	33.53	20.96	40.15	32.05	46.01	29.28	32.71
GBi-Net [23]	61.42	79.77	**67.69**	51.81	61.25	60.37	55.87	60.67	53.89	37.32	**29.77**	42.41	36.30	47.69	31.11	36.93
EPP-MVSNet [22]	61.68	77.86	60.54	52.96	62.33	61.69	60.34	62.44	55.30	35.72	21.28	39.74	35.34	49.21	30.00	**38.75**
TransMVSNet [8]	63.52	80.92	65.83	56.94	62.54	63.06	60.00	60.20	**58.67**	37.00	24.84	44.59	34.77	46.49	**34.69**	36.62
Uni-MVSNet [25]	64.36	81.20	66.43	53.11	63.46	66.09	64.84	62.23	57.53	38.96	28.33	44.36	39.74	52.89	33.80	34.63
RA-MVSNet (ours)	**65.72**	**82.44**	66.61	**58.40**	64.78	**67.14**	**65.60**	62.74	58.08	**39.93**	29.17	**46.05**	**40.23**	**53.22**	34.62	36.30

Table 1. **Quantitative results of F-score on Tanks and Temples benchmark.** The best results in each category are in **bold**. "Mean" refers to the mean F-score of all scenes (higher is better). Our RA-MVSNet achieves competitive results on both intermediate and advanced set.

Thus, the error range of case (b) and (c) can be expressed by the following equation

$$\mathbf{e}_b^2 \leq \min\left(d\left(\mathbf{p_j'}, \mathbf{o'}\right)^2, d\left(\mathbf{p_k'}, \mathbf{o'}\right)^2\right) \leq \frac{d\left(\mathbf{p_j'}, \mathbf{p_k'}\right)^2}{4},$$

$$\mathbf{e}_c^2 \leq \min\left(d\left(\mathbf{p_j'}, \mathbf{o''}\right)^2, d\left(\mathbf{p_k'}, \mathbf{o''}\right)^2, d\left(\mathbf{p_l'}, \mathbf{o''}\right)^2\right)$$
$$\leq \frac{\min\left(d(\mathbf{p_j'}, \mathbf{p_k'})^2, d(\mathbf{p_k'}, \mathbf{p_l'})^2, d(\mathbf{p_j'}, \mathbf{p_l'})^2\right)}{3} \quad (11)$$

where e_b^2 and e_c^2 are the square of error in case (b), (c), respectively. By combining these three cases covering all possible situations, we obtain the final error bound for the query point $\mathbf{p}$ as follows

$$0 \leq \mathbf{e}^2 = \max(\mathbf{e}_a^2, \mathbf{e}_b^2, \mathbf{e}_c^2) < d(\mathbf{p_j'}, \mathbf{p_{j+1}'})^2, \quad (12)$$

where $\mathbf{e}$ is the general error of the query point $\mathbf{p}$. $\mathbf{p_j'}$ and $\mathbf{p_{j+1}'}$ are the two adjacent surface points. This inequality shows that the square of error $\mathbf{e}^2$ does not exceed the square of distance between the two points that are reprojected from two adjacent pixels.

4. Experiments

In this section, we conduct the experiments and ablation studies on MVS benchmark datasets. The experimental results results show that our proposed RA-MVSNet approach achieves the start-of-the-art performance.

4.1. Implementation Setup

4.1.1 Training.

Like the previous methods [25, 40], our proposed RA-MVSNet is trained on DTU dataset for DTU evaluation, which is finetuned on BlendedMVS dataset for Tanks and Temples benchmark. As for DTU dataset, we use 79 scenes for training, 18 scenes for validation and the rest of data for evaluation. The original image size is 1200×1600, and each scene have 7 different lighting conditions. We crop the rectified images into 512×640. Meanwhile, we utilize the finer DTU ground truth as [32]. Similar to [13], we implement our RA-MVSNet in three stages with $\frac{1}{4}$, $\frac{1}{2}$ and original input images, respectively. From low-resolution to high-resolution stages, the number of depth hypothesis is 64, 32 and 8. Their corresponding depth intervals are set to 4, 2 and 1. In the training, the number of input images N is set to 5. Thus, there are single reference image and four source images. Our model is trained for 16 epochs with Adam optimizer [16]. The initial learning rate is 0.001, which is multiplied by 0.5 after 10, 12 and 14 epochs. Since the premature introduction of 3D CNN for SDF prediction may lead to slow convergence, we start the training of this branch at 10 epoch. The fusion parameter θ is set to 0.1. As for BlendedMVS dataset, we train for 10 epochs with an initial learning rate of 0.0002, which is down-scaled by a factor of 2 after 6 and 8 epochs. During finetuning, the number of input images is 10 with the original size of 576 $\times$ 768. The batch size is 2 on two NVIDIA RTX 2080Ti for DTU dataset, which is set to one on single NVIDIA Tesla P40 for BlendedMVS dataset.

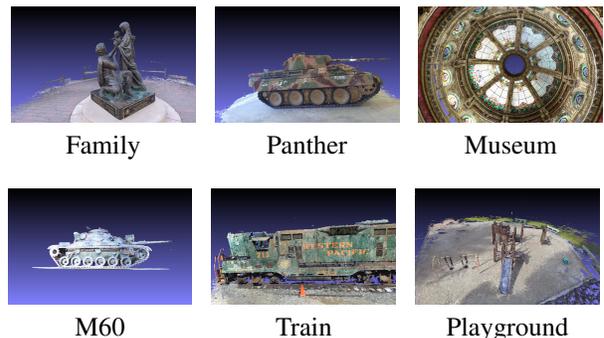|		
Family	Panther	Museum
M60	Train	Playground

Figure 7. **Qualitative results on T&T dataset.** Our RA-MVSNet still performs well in large outdoor scenes with complex lighting.

4.1.2 Testing.

When testing on the DTU dataset, the resolution is 864 × 1152, and the number of input images N is set to 5. Besides, we set the number of hypothetical planes for the three stages to 64, 32, and 8, which are the same as training. As for Tanks and Temples dataset, the resolution of input images is either 1024 × 1920 or 1024× 2048. The number of input images is 11 like [40]. To evaluate on the DTU dataset and the Tanks and Temples dataset, we use NVIDIA Tesla P40 GPU with 24G RAM. For the results on DTU, we report the evaluation metrics (accuracy, completeness and overall) described in [1]. For the benchmark results on Tanks and Temples, we report the F-score metric.

4.2. Results on Tanks and Temples

Our RA-MVSNet maintains the promising performance in large-scale, complex lighting scenes, which achieves the best score on Tanks and Temples dataset. Similar to the previous method [37], we employ the dynamic geometric consistency strategies.

For fair evaluation, we compare our method against current excellent work. The corresponding quantitative results on intermediate and advanced sets are reported in Table 1. It can be clearly seen that our method achieves the state-of-the-art performance on both intermediate and advanced sets. Specifically, RA-MVSNet obtains the best F-score of **65**.**72** and **39**.**93** (higher is better) on intermediate and advanced subset, respectively. Moreover, it performs the best on 5 scenes and 3 scenes of two subset, respectively. Notably, intermediate subset mostly contains one object to be reconstructed, and the advanced subset has the large-scale outdoor scenes. Our method achieves the best performance on both subsets. This shows that our model is effective in various scenarios. Moreover, Fig. 7 gives some qualitative results on two subsets, which demonstrate that our model exhibits the strongest generalization and robustness in textureless and object boundary regions.

Method	ACC.(mm) ↓	Comp.(mm) ↓	Overall(mm) ↓
Furu [11]	0.613	0.941	0.777
Gipuma [12]	**0.283**	0.873	0.578
COLMAP [27]	0.400	0.664	0.532
SurfaceNet [15]	0.450	1.040	0.745
MVSNet [40]	0.396	0.527	0.462
Point-MVSNet [4]	0.342	0.411	0.376
AA-RMVSNet [32]	0.376	0.339	0.357
CasMVSNet [13]	0.325	0.385	0.355
UCS-Net [6]	0.338	0.349	0.344
Uni-MVSNet [25]	0.352	0.278	0.315
TransMVSNet [8]	0.321	0.289	0.305
GBi-Net [23]	0.327	**0.268**	0.298
RA-MVSNet (ours)	0.326	**0.268**	**0.297**

Table 2. **Quantitative results on DTU evaluation set.** The best results in each category are in **bold**. Our model ranks the first in terms of Completeness and Overall metrics.

4.3. Results on DTU

As in [13, 40, 42], we make use of geometric and photometric constraints for filtering. Moreover, we employ the fusion method in Gipuma [12] similar to [13,40,42]. The final results are evaluated on DTU testing set by two metrics, accuracy and completeness. We compare our RA-MVSNet with previous methods. The quantitative results are summarized in Table 2. It can be observed that our RA-MVSNet outperforms both traditional methods and learning-based approaches. For the accuracy, the traditional method [12] achieves the best results. For the completeness metric, our method achieves the state-of-the-art performance. Overall, our RA-MVSNet method ranks the first, which achieves **0**.**297** score. Fig. 8 shows some qualitative results on DTU testing dataset compared against other methods. It can be seen that RA-MVSNet obtains more complete reconstruction results with less outlier.

4.4. Ablation Studies

As metioned above, we introduce the signed distance prediction branch in MVS network, which can not only improve the completeness of reconstruction results but also generate the explicit mesh surfaces. In the following, we show the effectiveness of the introduced branch. CasMVS-Net [13] is treated as baseline, which only has the depth prediction branch at each stage. Besides, we explore the effect of fusion parameter θ on performance.

4.4.1 SDF Branch

As shown in Table 3, we evaluate the performance of baseline and the model with SDF supervised branch. From the results, it can be obviously seen that the SDF branches are able to improve the performance, especially on the integrity of reconstruction results. This is because the approximate signed distance introduces the extra supervision on depth map prediction. The comparison experiment of whether to fuse these two branches shows that the introduction of SDF branch fusion can effectively improve the accuracy. This is because the SDF branch can remove the outliers whose sign distance exceeds the threshold. Meanwhile, our RA-MVSNet is able to generate both point clouds and mesh using two branches, respectively. However, the baseline method only employs point clouds as representation.

4.4.2 Local Patch Size

Instead of using global search, it is very efficient to compute the ground truth of distance volume in training by local search. Obviously, the size of patch affects the accuracy which affects the performance of RA-MVSNet. Therefore, we examine the performance of the model with different patch sizes, as shown in Fig. 9. When k becomes large,
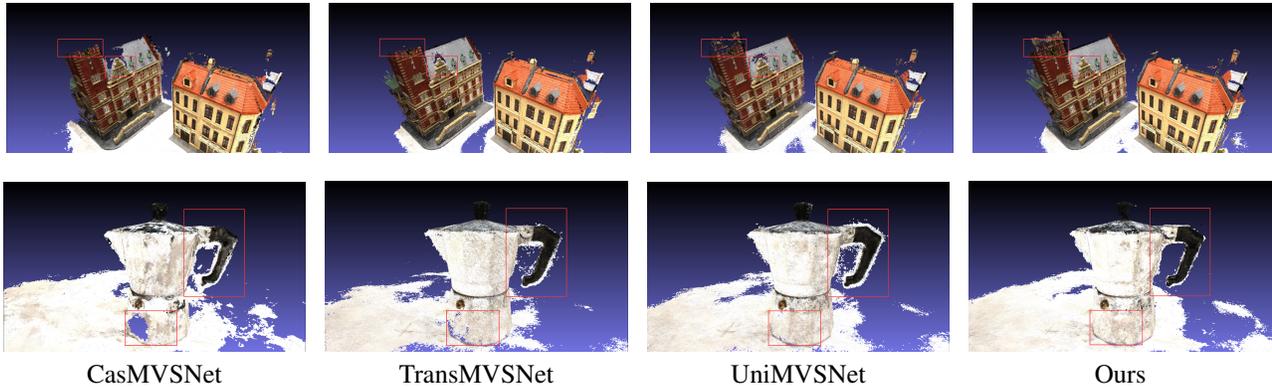
| | CasMVSNet | TransMVSNet | UniMVSNet | Ours |

Figure 8. **Qualitative results on DTU dataset.** Our RA-MVSNet produces more complete and less outlier results than the previous methods like Uni-MVSNet [25] and TransMVSNet [8].

Method	Branch		Representation		DTU			Tanks and Temples		
	depth	SDF	point clouds	mesh	ACC.(mm)	Comp.(mm)	Overall(mm)	Prec.	Rec.	F-score
Baseline	✓		✓		0.348	0.290	0.319	56.62	75.35	64.02
Two-branch(W/O fusion)	✓	✓	✓	✓	0.357	**0.262**	0.310	56.17	**77.65**	64.62
Two-branch(With fusion)	✓	✓	✓	✓	0.330	0.274	0.302	57.58	77.21	65.39
Two-branch(With fusion) + RFP	✓	✓	✓	✓	**0.326**	0.268	**0.297**	**58.68**	75.23	**65.72**

Table 3. **Ablation study on DTU and T&T evaluation set.** "RFP" refer to Recursive Feature Pyramid for feature extraction. The Baseline is the original CasMVSNet [13]. Our RA-MVSNet with fusion of two branches outperforms in the overall metric.

(a) Point cloud metric on DTU

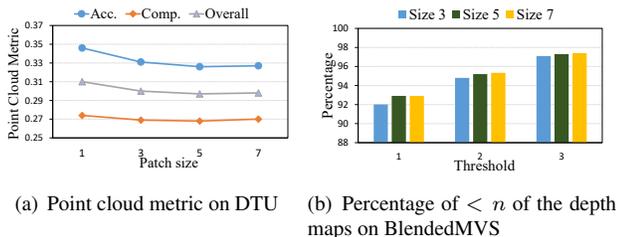(b) Percentage of $< n$ of the depth maps on BlendedMVS

Figure 9. **Ablation studies on patch size.** We use depth metrics and point cloud metrics to evaluate on BlendedMVS and DTU.

the performance of patch-based local search gradually increases. Moreover, the improvement is saturated after k exceeds 5. Therefore, we finally choose $k = 5$ as the patch size.

4.4.3 Fusion Threshold θ

The threshold θ is the parameter that trades off the fusion of two branches. From the essence of signed distance, the point closer to the surface has a smaller signed distance value. Therefore, we finally set θ to 0.1 as the threshold in this paper. As shown in Table 4, we compare the performance of different models with various thresholds. The model without fusing two branches is treated as the baseline. The experimental results show that the reconstruction accuracy gradually decreases with the larger threshold θ while the completeness gradually increases.

Method	ACC.(mm)	Comp.(mm)	Overall(mm)
Baseline(W/O fusion)	0.357	**0.262**	0.310
RA-MVSNet($\theta = 0.1$)	**0.326**	0.268	**0.297**
RA-MVSNet($\theta = 0.2$)	0.338	0.266	0.302
RA-MVSNet($\theta = 0.5$)	0.355	0.265	0.310
RA-MVSNet($\theta = 1.0$)	0.357	**0.262**	0.310

Table 4. **Ablation study on fusion threshold.**

5. Conclusion

In this paper, we proposed a novel RA-MVSNet approach to recover the detailed 3D scenes by taking advantage of cost volume using both depth and SDF branches. The SDF supervision enabled more hypothetical planes for the depth prediction, especially in textureless and boundary regions. Furthermore, the sided distance was employed to represent the ground truth signed distance for training, which can be computed efficiently. Our proposed RA-MVSNet approach achieves the promising results on several challenging datasets, which outperforms the state-of-the-art methods.

In the future, we plan to only use the SDF branch for MVS, which can effectively reduce memory consumption.

Acknowledgments

References

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *IJCV*, 120(2):153–168, 2016. 2, 7

[2] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM TOG*, 28(3):24, 2009. 1

[3] Neill DF Campbell, George Vogiatzis, Carlos Hernández, and Roberto Cipolla. Using multiple hypotheses to improve depth-maps for multi-view stereo. In *ECCV*, pages 766–779. Springer, 2008. 2

[4] Rui Chen, Songfang Han, Jing Xu, and Hao Su. Point-based multi-view stereo network. In *ICCV*, pages 1538–1547, 2019. 1, 7

[5] Rui Chen, Songfang Han, Jing Xu, and Hao Su. Visibility-aware point-based multi-view stereo network. *PAMI*, pages 1–1, 2020. 2, 6

[6] 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. In *CVPR*, pages 2524–2534, 2020. 2, 7

[7] Yikang Ding, Zhenyang Li, Dihe Huang, Zhiheng Li, and Kai Zhang. Enhancing multi-view stereo with contrastive matching and weighted focal loss. *arXiv preprint arXiv:2206.10360*, 2022. 2

[8] Yikang Ding, Wentao Yuan, Qingtian Zhu, Haotian Zhang, Xiangyue Liu, Yuanjiang Wang, and Xiao Liu. Transmvsnet: Global context-aware multi-view stereo network with transformers. In *CVPR*, pages 8585–8594, 2022. 2, 6, 7, 8

[9] Pascal Fua and Yvan G Leclerc. Object-centered surface reconstruction: Combining multi-image stereo and shading. *IJCV*, 16(1):35–56, 1995. 2

[10] Clement Fuji Tsang, Maria Shugrina, Jean Francois Lafleche, Towaki Takikawa, Jiehan Wang, Charles Loop, Wenzheng Chen, Krishna Murthy Jatavallabhula, Edward Smith, Artem Rozantsev, Or Perel, Tianchang Shen, Jun Gao, Sanja Fidler, Gavriel State, Jason Gorski, Tommy Xiang, Jianing Li, Michael Li, and Rev Lebaredian. Kaolin: A pytorch library for accelerating 3d deep learning research. https://github.com/NVIDIAGameWorks/kaolin, 2022. 4

[11] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE TPAMI*, 32(8):1362–1376, 2009. 1, 2, 7

[12] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *ICCV*, pages 873–881, 2015. 2, 7

[13] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *CVPR*, pages 2495–2504, 2020. 2, 3, 6, 7, 8

[14] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *CVPR*, pages 2821–2830, 2018. 1

[15] Mengqi Ji, Juergen Gall, Haitian Zheng, Yebin Liu, and Lu Fang. Surfacenet: An end-to-end 3d neural network for multiview stereopsis. In *ICCV*, pages 2307–2315, 2017. 1, 7

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[17] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM TOG*, 36(4):1–13, 2017. 2

[18] Kiriakos N Kutulakos and Steven M Seitz. A theory of shape by space carving. *IJCV*, 38(3):199–218, 2000. 2

[19] Maxime Lhuillier and Long Quan. A quasi-dense approach to surface reconstruction from uncalibrated images. *IEEE TPAMI*, 27(3):418–433, 2005. 2

[20] Keyang Luo, Tao Guan, Lili Ju, Haipeng Huang, and Yawei Luo. P-mvsnet: Learning patch-wise matching confidence aggregation for multi-view stereo. In *ICCV*, pages 10452–10461, 2019. 2

[21] Keyang Luo, Tao Guan, Lili Ju, Yuesong Wang, Zhuo Chen, and Yawei Luo. Attention-aware multi-view stereo. In *CVPR*, pages 1590–1599, 2020. 2

[22] Xinjun Ma, Yue Gong, Qirui Wang, Jingwei Huang, Lei Chen, and Fan Yu. Epp-mvsnet: Epipolar-assembling based depth prediction for multi-view stereo. In *ICCV*, pages 5732–5740, 2021. 2, 6

[23] Zhenxing Mi, Chang Di, and Dan Xu. Generalized binary search network for highly-efficient multi-view stereo. In *CVPR*, 2022. 2, 6, 7

[24] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *CVPR*, June 2019. 3

[25] Rui Peng, Rongjie Wang, Zhenyu Wang, Yawen Lai, and Ronggang Wang. Rethinking depth estimation for multi-view stereo: A unified representation. In *CVPR*, 2022. 2, 6, 7, 8

[26] Siyuan Qiao, Liang-Chieh Chen, and Alan Yuille. Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. In *CVPR*, pages 10213–10224, June 2021. 3

[27] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, pages 501–518, 2016. 1, 2, 7

[28] Steven M Seitz and Charles R Dyer. Photorealistic scene reconstruction by voxel coloring. *IJCV*, 35(2):151–173, 1999. 2

[29] Christian Sormann, Patrick Knöbelreiter, Andreas Kuhn, Mattia Rossi, Thomas Pock, and Friedrich Fraundorfer. Bp-mvsnet: Belief-propagation-layers for multi-view-stereo. In *3DV*, pages 394–403. IEEE, 2020. 2

[30] Fangjinhua Wang, Silvano Galliani, Christoph Vogel, and Marc Pollefeys. Itermvs: Iterative probability estimation for efficient multi-view stereo. In *CVPR*, pages 8606–8615, June 2022. 2

[31] Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Pablo Speciale, and Marc Pollefeys. Patchmatchnet: Learned multi-view patchmatch stereo. In *CVPR*, pages 14194–14203, 2021. 2, 6

[32] Zizhuang Wei, Qingtian Zhu, Chen Min, Yisong Chen, and Guoping Wang. Aa-rmvsnet: Adaptive aggregation recurrent multi-view stereo network. In *ICCV*, 2021. 2, 6, 7

[33] Junhua Xi, Yifei Shi, Yijie Wang, Yulan Guo, and Kai Xu. Raymvsnet: Learning ray-based 1d implicit fields for accurate multi-view stereo. In *CVPR*, pages 8595–8605, June 2022. 2, 6

[34] Qingshan Xu, Martin R. Oswald, Wenbing Tao, Marc Pollefeys, and Zhaopeng Cui. Non-local recurrent regularization networks for multi-view stereo, 2021. 2

[35] Qingshan Xu and Wenbing Tao. Multi-scale geometric consistency guided multi-view stereo. In *CVPR*, pages 5483–5492, 2019. 2

[36] Qingshan Xu and Wenbing Tao. Learning inverse depth regression for multi-view stereo with correlation cost volume. In *AAAI*, volume 34, pages 12508–12515, 2020. 2

[37] Jianfeng Yan, Zizhuang Wei, Hongwei Yi, Mingyu Ding, Runze Zhang, Yisong Chen, Guoping Wang, and Yu-Wing Tai. Dense hybrid recurrent multi-view stereo net with dynamic consistency checking. In *ECCV*, pages 674–689, 2020. 2, 6, 7

[38] Jiayu Yang, Wei Mao, Jose M Alvarez, and Miaomiao Liu. Cost volume pyramid based depth inference for multi-view stereo. In *CVPR*, pages 4877–4886, 2020. 6

[39] Zhenpei Yang, Zhile Ren, Qi Shan, and Qixing Huang. Mvs2d: Efficient multi-view stereo via attention-driven 2d convolutions. In *CVPR*, pages 8574–8584, 2022. 2

[40] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *ECCV*, pages 767–783, 2018. 1, 2, 3, 6, 7

[41] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *CVPR*, pages 5525–5534, 2019. 2

[42] 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. In *CVPR*, pages 1790–1799, 2020. 7

[43] Anzhu Yu, Wenyue Guo, Bing Liu, Xin Chen, Xin Wang, Xuefeng Cao, and Bingchuan Jiang. Attention aware cost volume pyramid based multi-view stereo network for 3d reconstruction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 175:448–460, 2021. 2

[44] Zehao Yu and Shenghua Gao. Fast-mvsnet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement. In *CVPR*, pages 1949–1958, 2020. 2

[45] Jingyang Zhang, Yao Yao, Shiwei Li, Zixin Luo, and Tian Fang. Visibility-aware multi-view stereo network. In *BMVC*, 2020. 2

[46] Xudong Zhang, Yutao Hu, Haochen Wang, Xianbin Cao, and Baochang Zhang. Long-range attention network for multi-view stereo. In *WACV*, pages 3782–3791, 2021. 2