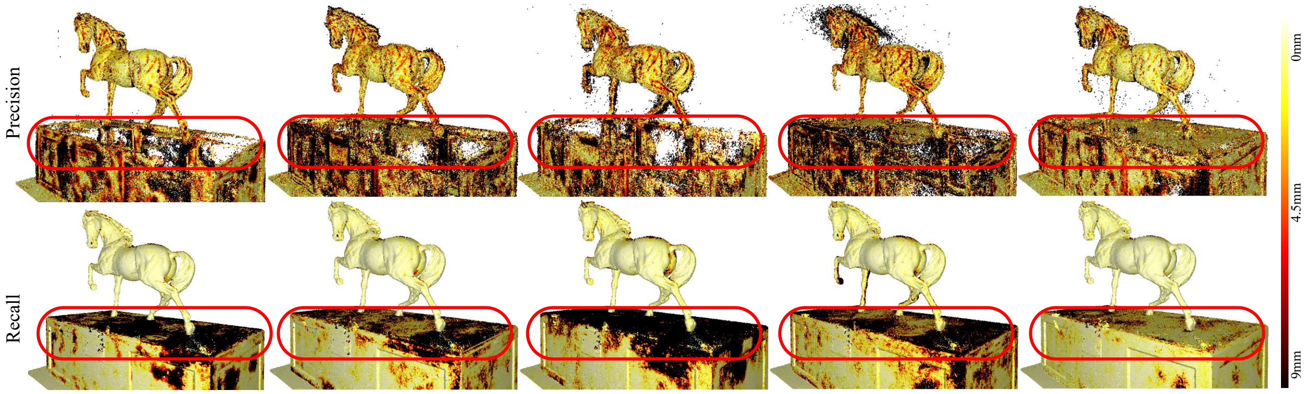
GoMVS: Geometrically Consistent Cost Aggregation for Multi-View Stereo

Jiang Wu[1] * Rui Li[1,2] * Haofei Xu[2,3] Wenxun Zhao[1] Yu Zhu[1] † Jinqiu Sun[1] Yanning Zhang[1] †
[1]Northwestern Polytechnical University [2]ETH Zürich [3]University of Tübingen, Tübingen AI Center



(a) MVSFormer [2] (b) RA-MVSNet [45] (c) ET-MVSNet [14] (d) GeoMVSNet [46] (e) Ours

Figure 1. **Comparison of reconstruction errors on Tanks and Temple benchmark.** We show precision and recall error maps for the "Horse" scan. Our method demonstrates notable improvements over existing methods in challenging areas.

Abstract

Matching cost aggregation plays a fundamental role in learning-based multi-view stereo networks. However, directly aggregating adjacent costs can lead to suboptimal results due to local geometric inconsistency. Related methods either seek selective aggregation or improve aggregated depth in the 2D space, both are unable to handle geometric inconsistency in the cost volume effectively. In this paper, we propose GoMVS to aggregate geometrically consistent costs, yielding better utilization of adjacent geometries. More specifically, we correspond and propagate adjacent costs to the reference pixel by leveraging the local geometric smoothness in conjunction with surface normals. We achieve this by the geometric consistent propagation (GCP) module. It computes the correspondence from the adjacent depth hypothesis space to the reference depth space using surface normals, then uses the correspondence to propagate adjacent costs to the reference geometry, followed by a convolution for aggregation. Our method achieves new state-of-the-art performance on DTU, Tanks & Temple, and ETH3D datasets. Notably, our method ranks 1st on the Tanks & Temple Advanced benchmark. Code is available at https://github.com/Wuuu3511/GoMVS.

1. Introduction

Multi-view stereo (MVS) is a fundamental computer vision problem that recovers 3D shapes from posed images by multi-view correspondence matching [21]. Recent learning-based MVS [11, 25, 30, 38] estimates scene depth from the cost volume computed by geometric matching, which delivers latent geometric cues crucial for the final depth [7]. However, the initial cost volume can suffer from challenging matching conditions, *e.g.*, varying illumination, textless areas, or repetitive patterns, leading to suboptimal pixel-wise costs that hamper accurate estimations.

To mitigate this issue, cost aggregation plays an important role in removing matching ambiguities and improving discriminativeness by using the neighboring information. However, the adjacent costs may deliver *inconsistent* depth cues due to the gradual changes in local geometry. As a result, the aggregated costs are not geometrically guaranteed to have the highest matching score at the real reference depth, leading to suboptimal depth predictions. The widely adopted cascade framework [7] can potentially exacerbate this issue as the adjacent costs can have more divergent costs due to the shifted depth hypotheses.

As the geometric inconsistency is a common challenge in multi-view stereo and 2-view stereo matching, related methods either adopt learned aggregation [25, 31] or enforce consistency to the aggregated depth [9, 15, 42]. Specif-

[*] indicates equal contributions and † indicates corresponding authors.

ically, some methods [25, 31] adopt adaptive aggregation schemes to allow networks to select pixels that potentially correlate well and contribute to the reference pixel's geometry. However, they heavily rely on network capabilities and do not guarantee geometric plausibility from the selected costs. Other methods [9, 19] seek to refine or regularize the aggregated depth values using jointly estimated surface normals. However, these methods only refine the output depth in 2D image space and are inherently unable to handle inconsistencies in the cost volume, which is vital for MVS methods.

In this paper, we propose GoMVS that aggregates geometrically consistent costs, allowing better utilization of adjacent geometries. Considering that the local geometry is usually smooth and exhibits gradual changes, we leverage the local smoothness to correspond and propagate adjacent costs to the reference cost. We achieve this by the geometrically consistent aggregation scheme, which operates on the local convolution window and propagates adjacent costs with the geometrically consistent propagation (GCP) module. The GCP module computes the correspondences from the adjacent cost's hypothesized depth space to the reference cost's depth space, using back-projected depth hypotheses and the surface normal. Then, it propagates the adjacent costs to the reference by interpolating cost scores with respect to the correspondence. After propagating adjacent costs within a local window, we aggregate them using standard convolutions. Unlike previous methods [9, 15, 19] that refine the predicted depth in the 2D space, our method incorporates geometric consistency in the cost space, yielding a better utilization of adjacent geometries. As surface normal is crucial for corresponding and propagating local costs, we further investigate different choices of normal predictions. We find appropriately applying off-the-shelf monocular normal models enables smooth and robust aggregation across datasets. We conduct extensive experiments to evaluate our method's effectiveness, and our method achieves new state-of-the-art on DTU, Tank & Temple, as well as the ETH3D dataset. Our contributions are summarized as follows:

- We propose GoMVS to aggregate geometrically consistent costs, allowing better utilization of adjacent geometries.
- We propose a geometrically consistent propagation (GCP) module that allows geometrically plausible correspondence and propagation in cost space.
- We investigate different choices of normal computation and find that properly applying the monocular surface normal model performs well across datasets.

2. Related Works

2.1. Learning-based MVS Methods

Multi-View Stereo (MVS) aims to reconstruct 3D scenes from multiple posed images. In recent years, learning-based methods have exhibited promising results. MVSNet [38] uses differentiable homography to construct the cost volume and employs a 3D U-Net for regularization. Subsequent works improve this framework in several ways. RNN-based methods [29, 35, 39]and coarse-to-fine approaches [3, 7, 17, 25, 36] reduce memory consumption through by designing efficient structures. Another group of methods [4, 12, 14] devises local or global attention modules to enhance input feature representations. MVSFormer [2] incorporates an additional pre-trained transformer network, enhancing the performance of MVS with a powerful feature extractor. However, it lacks further exploration in terms of geometry. GeoMVSNet [46] utilizes the coarse depth map to extract additional geometric features. In addition, [27, 34, 44] have designed pixel-wise visibility modules to handle occlusions.

2.2. Cost Volume Aggregation

As cost volume is vital for multi-view depth estimation, recent works introduce different cost aggregation methods to the depth network. NP-CVP-MVSNet [37] introduces sparse convolution to aggregate matching costs at the same depth range. WT-MVSNet [12] employs a cost transformer to generate a more complete and smoother probability volume. GeoMVSNet [46] incorporates the coarse probability volume to enhance the matching discriminative ability. While these methods improve the capability of regularization networks, the local geometric inconsistency of the cost volume still remains and poses challenges for the final aggregation results.

2.3. Normal Assisted Depth Estimation

Surface normal provides rich geometric details and has been widely applied in recent years to depth estimation tasks. Traditional MVS methods [20, 32, 33] optimize depth and normal hypotheses simultaneously by constructing a planar prior model. Inspired by traditional methods, SP-Net [28] performs slanted plane cost aggregation by learning parameterized local planes. NAPV-MVS [24] uses local normal similarity to emphasize the most relevant adjacent costs. NR-MVSNet [10] utilizes depth-normal consistency to adaptively expand the hypothesis range, providing broader matchings to assist depth inference. However, these methods do not address the local inconsistency issue. GeoNet [19] proposes a monocular depth estimation method that uses kernel regression to refine output depth with normals. However, it is sensitive to noisy outputs and is inherently incapable of handling cost volume inputs. An-

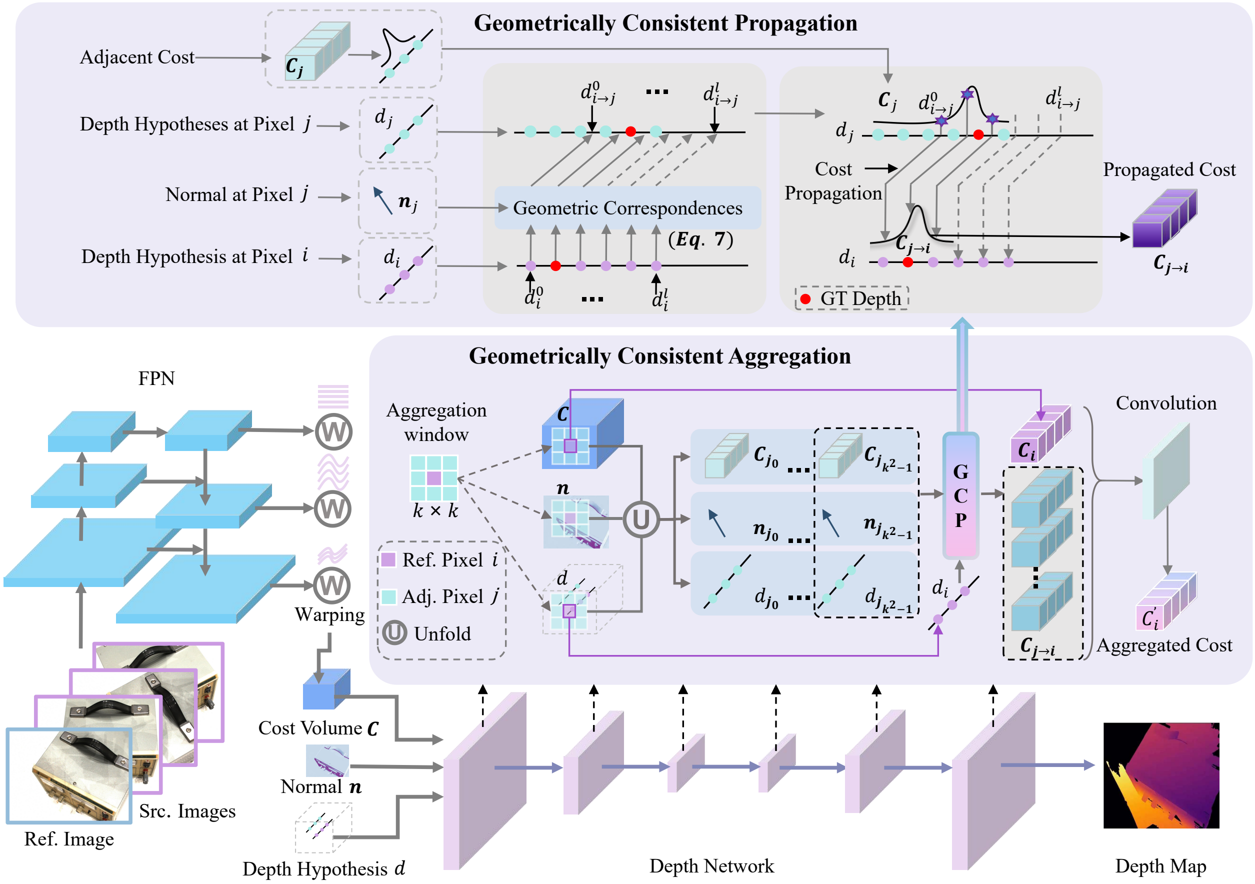


Figure 2. **Overview of our method.** Given a reference image and a set of source images, we use FPN to extract multi-scale features for cost volume reconstruction. To conduct geometrically consistent aggregation within the local window, we collect adjacent geometric cues and send them to the proposed geometrically consistent propagation (GCP) module, which computes the correspondence from the adjacent depth hypothesis space to the reference depth space. The resulting costs are endowed with geometric consistency, which facilitates better utilization of adjacent geometry and can be aggregated by the convolution.

other line of works [9, 15, 42] proposes the depth-normal consistency loss to enhance the network's perception of geometric cues. Unlike these methods, our method leverages the normal to yield geometrically consistent costs in the 3D space, yielding better utilization of adjacent costs.

3. Methodology

Given a reference image $\mathbf{I}_0 \in \mathbb{R}^{H\times W\times 3}$ and N source images $\{\mathbf{I}_i\}_{i=1}^N$, as well as camera intrinsic $\{\mathbf{K}_i\}_{i=0}^N$ and extrinsic parameters $\{[\mathbf{R}_{0\to i};\mathbf{t}_{0\to i}]\}_{i=1}^N$, our goal is to estimate the depth map of $\mathbf{I}_0$ from multiple posed images. Fig. 2 shows an overview of our method. We first utilize multi-scale image features to build the cost volume (Sec. 3.1). We then introduce the geometrically consistent aggregation scheme (Sec. 3.2), which consists of the blocks in the depth network. We then investigate different choices for obtaining surface normals (Sec. 3.3).

3.1. Cost Volume Construction

We first apply a Feature Pyramid Network [13] to extract multi-scale image features $\{\mathbf{F}_i^s\}_{i=0}^N \in \mathbb{R}^{\frac{H}{2^s}\times\frac{W}{2^s}\times M}$, where s is the scale factor. For simplicity, we omit the superscript of s below. To build the cost volume in each stage, we first sample depth hypotheses d for each pixel in a predefined depth range. Through differentiable homography, we can compute the corresponding position $\mathbf{p}'$ of the reference image's pixel $\mathbf{p}$ in the source image,

$$\mathbf{p}' = \mathbf{K}_i[\mathbf{R}_{0\to i}(\mathbf{K}_0^{-1}\mathbf{p}d) + \mathbf{t}_{0\to i}], \tag{1}$$

where $\mathbf{R}$ and $\mathbf{t}$ denote the rotation and translation parameters and $\mathbf{K}$ are the intrinsic matrix. Let $\mathbf{F}(\mathbf{p})$ represents the feature vector at pixel $\mathbf{p}$, then the two-view feature correlation volume $\mathbf{V}$ at pixel $\mathbf{p}$ can be represented as

$$\mathbf{V}_i(\mathbf{p}) = \mathbf{F}_0(\mathbf{p})\cdot\mathbf{F}_i(\mathbf{p}'), \tag{2}$$

where $\cdot$ refers to the dot product. To aggregate multiple pair-wise cost volumes, we utilize a shallow network [25] to learn the pixel-wise weight maps $\mathbf{W}$. The weight computation takes place exclusively in the initial stage, while weight maps for subsequent stages are derived through upsampling from the previous stage. Then the multi-view aggregated cost volume $\mathbf{C}$ can be represented as:

$$\mathbf{C} = \frac{\sum_{i=1}^{N} \mathbf{W}_i \odot \mathbf{V}_i}{\sum_{i=1}^{N} \mathbf{W}_i}. \tag{3}$$

3.2. Geometrically Consistent Aggregation

An essential idea of cost aggregation is to leverage neighboring information to improve the discriminativeness of the cost volume, where the key is to find the most relevant neighbors and effectively aggregate their matching costs. To achieve this, typical convolution-based methods are limited to the size of the convolution kernel (*e.g.*, $3 \times 3 \times 3$), and geometric inconsistency is very likely to happen in this local region due to non-constant depth distributions within this kernel. It's also computationally inefficient to directly increase the kernel size to get improved performance.

In this paper, we observe in a small local region, many scenes can be approximated with a plane, which frequently exists in real-world scenarios. To this end, we propose to leverage this locally approximated planar structure to guide the cost aggregation process in a geometrically consistent manner. There exists an analytic relationship between the reference pixel's depth and its local neighbors, which can be leveraged to obtain more reliable cost candidates. Specifically, for each reference pixel, we first collect the geometric clues of its $k \times k$ spatial window to compute the correspondences of the depth hypothesis. Depending on the corresponding location, we propagate the adjacent costs to the reference pixel's depth space. Finally, we use a convolution layer to aggregate the propagated costs.

3.2.1 Local Geometric Clues Collection

We first collect local depth hypotheses and normal maps for each pixel within a spatial window. Specifically, given the depth hypotheses of shape $L \times H \times W$ and the normal map of shape $3 \times H \times W$, where L is the depth hypothesis number and H, W denotes the spatial dimension, we unfold each pixel with a $k \times k$ spatial window, yielding local intermediate depth hypotheses volume and normal map of shape $k^2 \times L \times H \times W$ and $k^2 \times 3 \times H \times W$, respectively. We then compute the depth hypothesis correspondences based on these intermediate geometric clues.

3.2.2 Geometrically Consistent Propagation

To better aggregate the high-quality costs of the adjacent pixels, we align the adjacent pixels' depth hypothesis to the depth space of the reference pixel. Based on depth correspondence, we perform geometrically consistent cost propagation (GCP). Firstly, we introduce the depth relationship among pixels within the same plane. Given a pixel's image coordinates (u, v) and depth $d(u, v)$, its 3D point $X(u, v)$ in the camera coordinate system can be represented as

$$X(u,v) = \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} \frac{u-c_x}{f_x} \\ \frac{v-c_y}{f_y} \\ 1 \end{bmatrix} d(u,v), \tag{4}$$

where c_x, c_y, f_x, and f_y are the parameters of camera intrinsic $\mathbf{K}$. For the given reference pixel i and adjacent pixel j, we model the relationship between $X(u_i, v_i)$ and $X(u_j, v_j)$ by leveraging local planar assumption and the surface normal $\mathbf{n}$. They satisfy the equation of

$$\mathbf{n}^\top (X(u_i, v_i) - X(u_j, v_j)) = 0. \tag{5}$$

According to Eq. (4) and Eq. (5) , the depth relationship between the reference pixel i and the adjacent pixels j can be represented as:

$$\frac{d(u_j, v_j)}{d(u_i, v_i)} = \frac{\mathbf{n}^\top \begin{bmatrix} \frac{u_i - c_x}{f_x} & \frac{v_i - c_y}{f_y} & 1 \end{bmatrix}^\top}{\mathbf{n}^\top \begin{bmatrix} \frac{u_j - c_x}{f_x} & \frac{v_j - c_y}{f_y} & 1 \end{bmatrix}^\top}. \tag{6}$$

We use $r_{ji} = \frac{d(u_j, v_j)}{d(u_i, v_i)}$ to denote the depth ratio between j and i, which describes the linear transformation of depth within the plane. Based on this, we can compute the depth hypothesis correspondences. Specifically, define $[d_i^1, ..., d_i^L]$ as the depth hypothesis in the pixel i's depth space, where L refers to the number of depth sampling levels. Each depth hypothesis is then mapped to pixel j's depth space through the depth ratio r_{ji}.

$$[d_{i\to j}^1, ..., d_{i\to j}^L] = [r_{ji} \times d_i^1, ..., r_{ji} \times d_i^L], \tag{7}$$

where $d_{i\to j}$ represents the mapping depth of pixel i's depth hypothesis in pixel j's depth space. We then propagate the matching cost of pixel j at the $d_{i\to j}$ to d_i. Let $\mathbf{C}_j$ denote the cost for pixel j. The propagated matching cost $\mathbf{C}_{j\to i}$ can be expressed as:

$$\mathbf{C}_{j\to i}(d_i^0, ..., d_i^l) = \mathbf{C}_j(d_{i\to j}^0, ..., d_{i\to j}^l). \tag{8}$$

Since depth hypotheses are discretely sampled at regular depth intervals within the depth range, we can conveniently use linear interpolation to implement the above process. With the definition $d_{i\to j}^m = d_j^n$, $\mathbf{C}_{j\to i}(d_i^m)$ can be expressed as:

$$\mathbf{C}_{j\to i}(d_i^m) = (\mathbf{C}_j(d_j^{\lceil n \rceil}) - \mathbf{C}_j(d_j^{\lfloor n \rfloor})) \frac{n - \lfloor n \rfloor}{\lceil n \rceil - \lfloor n \rfloor}. \tag{9}$$

We refer to this process as geometrically consistent propagation from j to i. It can generate geometrically consistent cost candidates for each reference pixel. Due to varying depth relationships between each pixel and its adjacent pixels, cost propagation generates an intermediate cost of $k^2M \times L \times H \times W$, where M is the channel dimension.

3.2.3 Aggregating Propagated Costs

Since the intermediate costs include $k \times k$ spatial information in the channel dimension, we thus aggregate the costs using convolutions with a kernel size of $1 \times 1 \times k$ and an expanded channel dimension k^2M, leading to the same parameters as the generic 3D convolutions with kernel size $k \times k \times k$.

We encapsulate GCP and the convolution into one geometrically consistent aggregation operator used to build the depth network. In particular, we still keep the 3D U-Net architecture proposed by MVSNet [38], while replacing each standard 3D convolution block with our proposed geometrically consistent aggregation operator. For the upsampling layer in the U-Net structure, we use the pixel shuffle to reorganize features and obtain a high-resolution cost volume.

3.3. Extracting Normal Cues

Since our approach uses the surface normal for cost aggregation, in this section, we study different methods for obtaining surface normals. We conduct experiments to demonstrate the effectiveness of each method in Sec. 4.4.

Depth to normal. The surface normal can be directly computed from the estimated depth. Since we use a three-stage cascade structure, we leverage the depth map from the g stage to generate the surface normal for the $g+1$ stage. The normal $\mathbf{n}$ can be computed [19] in closed form as:

$$\mathbf{n} = \frac{(\mathbf{A}^\intercal\mathbf{A})^{-1}\mathbf{A}^\intercal\mathbf{1}}{\|(\mathbf{A}^\intercal\mathbf{A})^{-1}\mathbf{A}^\intercal\mathbf{1}\|}, \tag{10}$$

where $\mathbf{A}$ is a matrix composed of the coordinates of all pixels within the local window. In addition to using estimated depth maps, we also compute the GT normal from the GT depth maps following the same protocol and use it to train our method for evaluating performances.

Cost to normal. In addition, inspired by [9], we use an additional network branch to directly regress the normal map from the cost volume in each stage, which is then used as a prior for geometrically consistent aggregation.

Off-the-shelf monocular surface normal. Monocular networks directly perceive surface geometry from deep features and can estimate reasonable solutions in regions with multi-view consistency ambiguities, which complements the task of MVS. Therefore, we explore an existing monocular normal estimation network Omnidata [5] to generate the surface normal. Since Omnidata is trained on low-resolution images, its normal prediction might become unreliable when the testing input resolution is increased. To tackle this, we adopt a divide-and-conquer approach following MonoSDF [43] to generate high-resolution normal cues. Specifically, we first divide the high-resolution image into multiple overlapping patches. Surface normal estimation is then independently conducted for each patch. Subsequently, the surface normal results are aligned and fused to generate a high-resolution normal map.

3.4. Optimization

We treat the MVS task as a classification problem and employ the winner-takes-all strategy to obtain the final depth map [39]. We use the cross-entropy loss (Eq. 11) in each stage, which is applied to the probability volume P and the ground truth one-hot volume $P^{'}$. Following [17], all depth out-of-range will be masked during the training stage.

$$\mathcal{L} = \sum_{i=1}^{d} -P_i^{'}\log(P_i). \tag{11}$$

4. Experiments

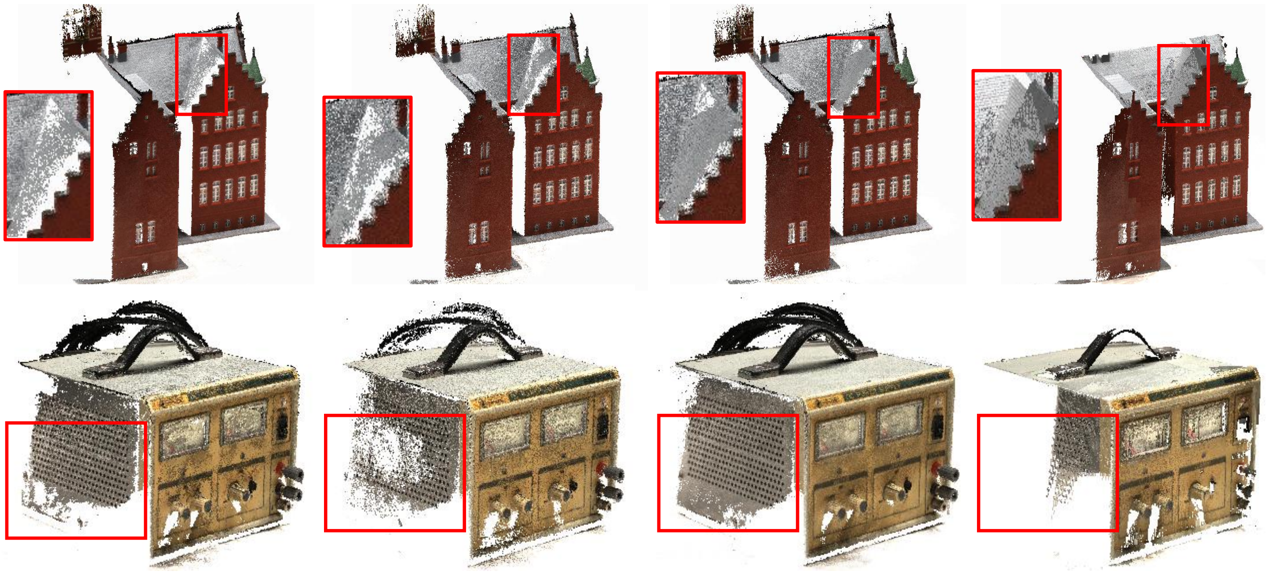
In this section, we evaluate our method on the DTU [1], ETH3D [22], and Tanks and Temple [8] datasets, respectively. Furthermore, we conducted multiple ablation experiments on the DTU dataset to validate the effectiveness of our method.

4.1. Datasets

DTU [1] dataset comprises 128 scenes in controlled laboratory environments, with models captured using structured light scanners. Each scene was scanned from the same 49 or 64 camera positions under 7 different lighting conditions. The official evaluation assesses the point cloud using distance metrics of accuracy and completeness. BlendedMVS [40] is a large-scale MVS dataset that consists of over 17,000 high-resolution images covering a variety of scenes, including urban environments, architecture, sculptures, and small objects. Tanks and Temples (TNT) [8] is a real-world dataset, divided into two sets, including 8 scenes in the intermediate set and 6 scenes in the advanced set. ETH3D [22] dataset consists of multiple indoor and outdoor scenes with large viewpoint variations. The quality of point clouds on the ETH3D and TNT datasets is measured using the percentage of precision and recall.

4.2. Implementation Details

Training Following the data partitioning of MVSNet, we first train the model on the DTU training set. Our network employs a three-stage cascade structure, with depth sampling at 48, 32, and 8 in each stage and depth intervals of



(a) TransMVSNet [4]

(b) GeoMVSNet [46]

(c) **Ours**

(d) GT

Figure 3. **Comparison of reconstruction results.** Our method reconstructs more complete results in challenging areas.

4, 1, and 0.5, respectively. We train our model with 5 input images, each having a resolution of 512×640. The model is optimized using Adam for 12 epochs, starting with an initial learning rate of 0.001 which is reduced by 0.5 after the 6 and 8 epochs. We then fine-tune the model on the BlendedMVS dataset with 9 images at a resolution of 576×768 for evaluation on Tanks and Temples and ETH3D datasets. During fine-tuning, we reduce the depth sampling interval of the last stage by half of its original value.

Evaluation When testing on the DTU dataset, we use 5 images at a resolution of 864×1152 as input and employ the depth map filtering method following [46] to generate the final point cloud. For the tanks and temple dataset, we carried out tests using 11 images with a resolution of 960×1920. In terms of depth map fusion, we employ the widely adopted dynamic fusion strategy [35]. Moreover, we conducted tests on the ETH3D dataset using images with a size of 1152×1536 and the depth map fusion strategy is consistent with IterMVS [26].

4.3. Benchmark Performance

Evaluation on DTU dataset. We compare both traditional methods and deep learning-based approaches. The quantitative evaluation results for point cloud reconstruction are shown in Tab 1. Our method achieves SOTA completeness and overall performance. It is worth noting that our method shows obvious improvement in completeness compared to previous methods. This demonstrates that our method can better use adjacent costs to propagate local geometries, resulting in a more complete reconstruction. Fig. 3 shows a comparison of our point cloud results with previous SOTA methods. We have more detailed and complete reconstructions in the challenge areas.

Method	Acc. ↓	Comp. ↓	Overall↓
Gipuma [6]	**0.283**	0.873	0.578
COLMAP [21]	0.400	0.664	0.532
NAPV-MVS [24]	0.367	0.375	0.371
AA-RMVSNet [29]	0.376	0.339	0.357
Vis-MVSNet [44]	0.369	0.361	0.365
CasMVSNet [7]	0.325	0.385	0.355
UniMVSNet [18]	0.352	0.278	0.315
MVSTER [27]	0.350	0.276	0.313
TransMVSNet [4]	0.321	0.289	0.305
GbiNet* [17]	0.314	0.295	0.305
RA-MVSNet [45]	0.326	0.268	0.297
GeoMVSNet [46]	0.331	0.259	0.295
ET-MVSNet [14]	0.329	0.253	0.291
MVSformer [2]	0.327	0.251	0.289
GoMVS	0.347	**0.227**	**0.287**

Table 1. **Quantitative results on DTU [1].** Our method achieves the best completeness and overall score. Moreover, the completeness of our point cloud outperforms previous methods by large margins.

Evaluation on Tanks and Temples dataset. We validated the generalization of our model on the Tanks and Temples dataset, and the quantitative results are shown in Table . We achieved the best performance on both the intermediate and advanced sets. Moreover, we *ranked 1st among all submitted results on the advanced set of the TNT benchmark*, which contains more complex scenes. It demonstrates the strong robustness and generalization ability of our method. Fig. 4 shows point cloud results on intermediate and advanced sets. Our method achieves detailed and complete reconstructions across different indoor

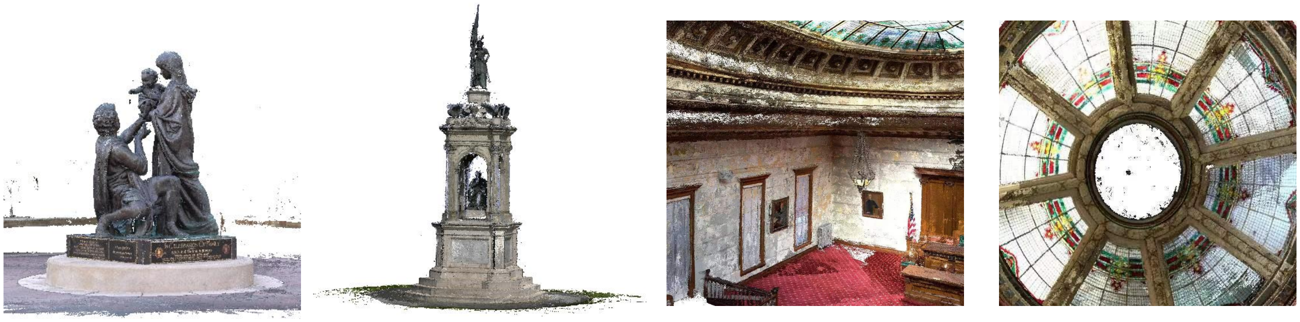

(a) Family (b) Francis (c) Courtroom (d) Musume

Figure 4. **Qualitative results on Tanks and Temples.** Our method achieves detailed and complete reconstructions across different scenes.

Methods	Intermediate									Advanced						
	Mean↑	Fam.	Fra.	Hor.	Lig.	M60	Pan.	Pla.	Tra.	Mean↑	Aud.	Bal.	Cou.	Mus.	Pal.	Tem.
COLMAP [21]	42.14	50.41	22.25	26.63	56.43	44.83	46.97	48.53	42.04	27.24	16.02	25.23	34.70	41.51	18.05	27.94
CasMVSNet [7]	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
Vis-MVSNet [44]	60.03	77.40	60.23	47.07	63.44	62.21	57.28	60.54	52.07	33.78	20.79	38.77	32.45	44.20	28.73	37.70
GBiNet [17]	61.42	79.77	67.69	51.81	61.25	60.37	55.87	60.67	53.89	37.32	29.77	42.12	36.30	47.69	31.11	36.93
EPP-MVSNet [16]	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 [4]	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
UniMVSNet [18]	64.36	81.20	66.43	53.11	64.36	66.09	64.84	62.23	57.53	38.96	28.33	44.36	39.74	52.89	33.80	34.63
D-MVSNet [41]	64.66	81.27	67.54	59.10	63.12	64.64	64.80	59.83	56.97	41.17	30.08	46.10	40.65	**53.53**	35.08	41.60
ET-MVSNet [14]	65.49	81.65	68.79	59.46	65.72	64.22	64.03	61.23	58.79	40.41	28.86	45.18	38.66	51.10	35.39	43.23
RA-MVSNet [45]	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
GeoMVSNet [46]	65.89	81.64	67.53	55.78	68.02	65.49	**67.19**	**63.27**	58.22	41.52	30.23	46.53	39.98	**5**3.05	35.98	43.34
MVSFormer [2]	66.37	82.06	**69.34**	60.49	**68.61**	65.67	64.08	61.23	59.33	40.87	28.22	46.75	39.30	**5**2.88	35.16	42.95
GoMVS	**66.44**	**82.68**	69.23	**69.19**	63.56	65.13	62.10	58.81	**60.80**	**43.07**	**35.52**	**47.15**	**42.52**	52.08	**36.34**	**44.82**

Table 2. **Quantitative results of F-score on Tanks and Temples benchmark.** Our method achieves the best F-score on both the "Intermediate" and the challenging "Advanced" set. Note that our method *ranks 1st on the official TNT Advanced Benchmark*.

Method	Acc. ↓	Comp. ↓	Overall ↓
Standard 3D-convolution [38]	0.365	0.265	0.315
Sparse convolution [37]	0.354	0.268	0.311
Spatial deformable aggregation [25]	0.363	0.257	0.310
Depth kernel regression [19]	0.369	0.262	0.316
Ours (GCA)	**0.347**	**0.227**	**0.287**

Table 3. **Comparison with different aggregation methods.** Our method significantly outperforms previous cost volume aggregation methods.

and outdoor scenes.

Evaluation on ETH3D dataset. The ETH3D dataset contains many challenging scenes, including scenes with textureless areas and large viewpoint variations. We compare our methods with previous methods and results are shown in Tab. 4. Our method achieves the best performance on both the validation set and the test split. In particular, it outperforms previous SOTA by a significant margin on the test split, demonstrating its generalization ability over existing methods.

4.4. Ablation Study

Comparison with different aggregation methods. To verify the effectiveness of utilizing adjacent geometry, we compare different cost aggregation and depth aggregation methods, and the results are shown in Tab. 3. Regarding the cost aggregation methods, Sparse convolution [37] aggregates the cost at the same depth without fully considering the depth geometry, resulting in certain improvements in performance compared with the baseline. PatchMatchNet [25] utilizes deformable convolutions to gather spatial matching costs and aggregate them using a lightweight 3D CNN. We replace the aggregation network with a 3D U-Net to ensure a fair comparison with the same parameter scale. PatchmatchNet heavily relies on network capabilities and does not guarantee geometric plausibility from the selected costs. As a result, it brings limited performance improvements (row #3).

Additionally, for the depth aggregation method, we refine the depth map on the baseline method by incorporating depth kernel regression proposed by GeoNet [19]. Using normal similarity to compute depth aggregation weights is prone to the influence of normal noise and cannot effectively utilize the abundant geometric information in the cost volume. This leads to a decline in the accuracy of the final point cloud (row #4). We utilize normal priors to guide cost aggregation, alleviating the challenge of geometric in-

Methods	Training			Test		
	Acc.↑	Comp. ↑	F-score↑	Acc.↑	Comp. ↑	F-score↑
COLMAP [21]	**91.85**	55.13	67.66	**91.97**	62.98	73.01
ACMM [32]	90.67	70.42	78.86	90.65	74.34	80.78
IterMVS [26]	73.62	61.87	66.36	76.91	72.65	74.29
GBi-Net [17]	73.17	69.21	70.78	80.02	75.65	78.40
MVSTER [27]	76.92	68.08	72.06	77.09	82.47	79.01
PVSNet [34]	83.00	71.76	76.57	81.55	83.97	82.62
EPP-MVSNet [16]	82.76	67.58	74.00	85.47	81.79	83.40
Vis-MVSNet [44]	83.32	65.53	72.77	86.86	80.92	83.46
EPNet [23]	79.36	**79.28**	79.08	80.37	**87.84**	83.72
GoMVS	81.22	77.65	**79.16**	85.50	86.85	**85.91**

Table 4. **Quantitative results on ETH3D dataset**. We show comparisons of reconstructed point clouds using percentage metric (%) at a threshold of 2cm. Our approach achieves the best performance with notable margins.

consistency and achieving the best performance among all aggregation methods.

Comparison with different depth receptive fields. Intuitively, 3D convolutions with larger receptive fields in the depth dimension can alleviate the cost inconsistency in the local range, by resorting to wider areas. Therefore, we compare our approach with variants directly expanding the depth receptive field. We keep the 3×3 spatial window size at each 3D convolution layer and experiment with kernel sizes of 3, 5, and 7 in the depth dimension on the baseline method. The quantitative results are shown in Tab. 5, we find that increasing the receptive field in the depth dimension leads to some certain improvement. However, due to the lack of geometric awareness, its performance is saturated when the dimension expands to a certain kernel size. In contrast, we use surface normal to geometrically guide the cost aggregation process. With a kernel size of only 3, our method achieves the best performance, outperforming other alternatives by clear margins.

Evaluation of different normal cues. Since the surface normal is important for guiding geometrically consistent aggregation, we further evaluate the effectiveness of different normal cues in Tab. 6. We first train and evaluate our method using the GT normal, which sets an upper bound for our method. As shown in the last row, it significantly improves the performance of point clouds, validating our method's effectiveness when using high-quality normal inputs. We further train and evaluate our method using depth-computed normals [19] or cost-computed normals [9], the results are suboptimal as they essentially rely on the quality of input depth, which can degrade in challenging areas. Though lacking multi-view consistency, monocular normals do not collapse in challenging geometric estimation regions of the cost volume. This reveals a nice property for monocular estimations. In addition to the DTU dataset, we also observe notable improvement using monocular surface normals on other benchmarks.

Aggregation kernel ($d \times h \times w$)	Acc. ↓	Comp. ↓	Overall ↓
Standard Conv3D($3 \times 3 \times 3$)	0.365	0.265	0.315
Standard Conv3D ($5 \times 3 \times 3$)	0.352	0.260	0.306
Standard Conv3D ($7 \times 3 \times 3$)	0.352	0.258	0.305
Proposed GCA ($3 \times 3 \times 3$)	**0.347**	**0.227**	**0.287**

Table 5. **Evaluation of aggregation receptive fields.** Directly expanding receptive fields along the depth dimension yields limited improvement and is easily saturated. In contrast, our method achieves the best performance with a kernel size of 3.

Method	Acc. ↓	Comp. ↓	Overall ↓
Ours + Depth-to-normal [19]	0.352	0.242	0.297
Ours + Cost-to-normal [9]	0.358	0.241	0.300
Ours + Mono-normal [5]	0.347	0.227	0.287
Ours + GT normal	**0.275**	**0.221**	**0.248**

Table 6. **Evaluation of different normal cues.** Our method with GT normal demonstrates remarkable performance (0.248). Among all estimated normals, the off-the-shelf monocular normal has the best performance.

5. Conclusion

In this paper, we propose GoMVS, which aggregates locally consistent geometries to better utilize adjacent geometry. By leveraging local smoothness in conjunction with surface normal, we propose geometrically consistent aggregation. it computes the correspondence from the adjacent depth hypotheses space to the reference depth space and propagates cost accordingly. Furthermore, we investigate different choices for generating normal priors and find that monocular cues effectively complement the MVS network. Our method achieves state-of-the-art performance on the DTU, Tanks and Temples, and ETH3D datasets.

Acknowledgements Y. Zhang was supported by NSFC (No.U19B2037) and the Natural Science Basic Research Program of Shaanxi (No.2021JCW-03). Y. Zhu was supported by NSFC (No.61901384). Y. Zhu, J. Sun, and Y. Zhang acknowledge the support from National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology.

References

[1] 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*, 120:153–168, 2016. 5, 6

[2] Chenjie Cao, Xinlin Ren, and Yanwei Fu. Mvsformer: Learning robust image representations via transformers and temperature-based depth for multi-view stereo. *arXiv preprint arXiv:2208.02541*, 2022. 1, 2, 6, 7

[3] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2524–2534, 2020. 2

[4] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8585–8594, 2022. 2, 6, 7

[5] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10786–10796, 2021. 5, 8

[6] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 873–881, 2015. 6

[7] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504, 2020. 1, 2, 6, 7

[8] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36 (4):1–13, 2017. 5

[9] Uday Kusupati, Shuo Cheng, Rui Chen, and Hao Su. Normal assisted stereo depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2189–2199, 2020. 1, 2, 3, 5, 8

[10] Jingliang Li, Zhengda Lu, Yiqun Wang, Jun Xiao, and Ying Wang. Nr-mvsnet: Learning multi-view stereo based on normal consistency and depth refinement. *IEEE Transactions on Image Processing*, 2023. 2

[11] Rui Li, Dong Gong, Wei Yin, Hao Chen, Yu Zhu, Kaixuan Wang, Xiaozhi Chen, Jinqiu Sun, and Yanning Zhang. Learning to fuse monocular and multi-view cues for multi-frame depth estimation in dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21539–21548, 2023. 1

[12] Jinli Liao, Yikang Ding, Yoli Shavit, Dihe Huang, Shihao Ren, Jia Guo, Wensen Feng, and Kai Zhang. Wt-mvsnet: window-based transformers for multi-view stereo. *Advances in Neural Information Processing Systems*, 35:8564–8576, 2022. 2

[13] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 3

[14] Tianqi Liu, Xinyi Ye, Weiyue Zhao, Zhiyu Pan, Min Shi, and Zhiguo Cao. When epipolar constraint meets non-local operators in multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18088–18097, 2023. 1, 2, 6, 7

[15] Xiaoxiao Long, Lingjie Liu, Christian Theobalt, and Wenping Wang. Occlusion-aware depth estimation with adaptive normal constraints. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 640–657. Springer, 2020. 1, 2, 3

[16] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5732–5740, 2021. 7, 8

[17] Zhenxing Mi, Chang Di, and Dan Xu. Generalized binary search network for highly-efficient multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12991–13000, 2022. 2, 5, 6, 7, 8

[18] 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*, pages 8645–8654, 2022. 6, 7

[19] Xiaojuan Qi, Renjie Liao, Zhengzhe Liu, Raquel Urtasun, and Jiaya Jia. Geonet: Geometric neural network for joint depth and surface normal estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 283–291, 2018. 2, 5, 7, 8

[20] Chunlin Ren, Qingshan Xu, Shikun Zhang, and Jiaqi Yang. Hierarchical prior mining for non-local multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3611–3620, 2023. 2

[21] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 501–518. Springer, 2016. 1, 6, 7, 8

[22] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 5

[23] Wanjuan Su and Wenbing Tao. Efficient edge-preserving multi-view stereo network for depth estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2348–2356, 2023. 8

[24] Wei Tong, Xiaorong Guan, Jian Kang, Poly ZH Sun, Rob Law, Pedram Ghamisi, and Edmond Q Wu. Normal assisted pixel-visibility learning with cost aggregation for multiview stereo. *IEEE Transactions on Intelligent Transportation Systems*, 23(12):24686–24697, 2022. 2, 6

[25] Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Pablo Speciale, and Marc Pollefeys. Patchmatchnet: Learned multi-view patchmatch stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14194–14203, 2021. 1, 2, 4, 7

[26] Fangjinhua Wang, Silvano Galliani, Christoph Vogel, and Marc Pollefeys. Itermvs: Iterative probability estimation for efficient multi-view stereo. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8606–8615, 2022. 6, 8

[27] Xiaofeng Wang, Zheng Zhu, Guan Huang, Fangbo Qin, Yun Ye, Yijia He, Xu Chi, and Xingang Wang. Mvster: epipolar transformer for efficient multi-view stereo. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*, pages 573–591. Springer, 2022. 2, 6, 8

[28] Yun Wang, Longguang Wang, Hanyun Wang, and Yulan Guo. Spnet: Learning stereo matching with slanted plane aggregation. *IEEE Robotics and Automation Letters*, 7(3): 6258–6265, 2022. 2

[29] Zizhuang Wei, Qingtian Zhu, Chen Min, Yisong Chen, and Guoping Wang. Aa-rmvsnet: Adaptive aggregation recurrent multi-view stereo network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6187–6196, 2021. 2, 6

[30] Jiang Wu, Rui Li, Yu Zhu, Wenxun Zhao, Jinqiu Sun, and Yanning Zhang. Boosting multi-view stereo with late cost aggregation. *arXiv preprint arXiv:2401.11751*, 2024. 1

[31] Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1959–1968, 2020. 1, 2

[32] Qingshan Xu and Wenbing Tao. Multi-scale geometric consistency guided multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5483–5492, 2019. 2, 8

[33] Qingshan Xu and Wenbing Tao. Planar prior assisted patchmatch multi-view stereo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12516–12523, 2020. 2

[34] Qingshan Xu, Wanjuan Su, Yuhang Qi, Wenbing Tao, and Marc Pollefeys. Learning inverse depth regression for pixelwise visibility-aware multi-view stereo networks. *International Journal of Computer Vision*, 130(8):2040–2059, 2022. 2, 8

[35] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV*, pages 674–689. Springer, 2020. 2, 6

[36] Jiayu Yang, Wei Mao, Jose M Alvarez, and Miaomiao Liu. Cost volume pyramid based depth inference for multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4877–4886, 2020. 2

[37] Jiayu Yang, Jose M Alvarez, and Miaomiao Liu. Non-parametric depth distribution modelling based depth inference for multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8626–8634, 2022. 2, 7

[38] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pages 767–783, 2018. 1, 2, 5, 7

[39] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5525–5534, 2019. 2, 5

[40] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1790–1799, 2020. 5

[41] Xinyi Ye, Weiyue Zhao, Tianqi Liu, Zihao Huang, Zhiguo Cao, and Xin Li. Constraining depth map geometry for multi-view stereo: A dual-depth approach with saddle-shaped depth cells. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 17661–17670, 2023. 7

[42] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5684–5693, 2019. 1, 3

[43] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *Advances in neural information processing systems*, 35:25018–25032, 2022. 5

[44] Jingyang Zhang, Shiwei Li, Zixin Luo, Tian Fang, and Yao Yao. Vis-mvsnet: Visibility-aware multi-view stereo network. *International Journal of Computer Vision*, 131(1): 199–214, 2023. 2, 6, 7, 8

[45] Yisu Zhang, Jianke Zhu, and Lixiang Lin. Multi-view stereo representation revist: Region-aware mvsnet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17376–17385, 2023. 1, 6, 7

[46] Zhe Zhang, Rui Peng, Yuxi Hu, and Ronggang Wang. Geomvsnet: Learning multi-view stereo with geometry perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21508–21518, 2023. 1, 2, 6, 7