GeoMVSNet: Learning Multi-View Stereo with Geometry Perception

Zhe Zhang\textsuperscript{1} Rui Peng\textsuperscript{1} Yuxi Hu\textsuperscript{2} Ronggang Wang\textsuperscript{1}
\textsuperscript{1}School of Electronic and Computer Engineering, Peking University, China
\textsuperscript{2}School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, China
doublez@stu.pku.edu.cn rgwang@pkusz.edu.cn

Abstract

Recent cascade Multi-View Stereo (MVS) methods can efficiently estimate high-resolution depth maps through narrowing hypothesis ranges. However, previous methods ignored the vital geometric information embedded in coarse stages, leading to vulnerable cost matching and sub-optimal reconstruction results. In this paper, we propose a geometry awareness model, termed GeoMVSNet, to explicitly integrate geometric clues implied in coarse stages for delicate depth estimation. In particular, we design a two-branch geometry fusion network to extract geometric priors from coarse estimations to enhance structural feature extraction at finer stages. Besides, we embed the coarse probability volumes, which encode valuable depth distribution attributes, into the lightweight regularization network to further strengthen depth-wise geometry intuition. Meanwhile, we apply the frequency domain filtering to mitigate the negative impact of the high-frequency regions and adopt the curriculum learning strategy to progressively boost the geometry integration of the model. To intensify the full-scene geometry perception of our model, we present the depth distribution similarity loss based on the Gaussian-Mixture Model assumption. Extensive experiments on DTU and Tanks and Temples (T&T) datasets demonstrate that our GeoMVSNet achieves state-of-the-art results and ranks first on the T&T-Advanced set. Code is available at https://github.com/doubleZ0108/GeoMVSNet.

1. Introduction

Multi-View Stereo (MVS) reconstructs the dense geometry representation of a scene from multiple overlapping photographs, which is an influential branch of three-dimensional (3D) computer vision and has been extensively studied for decades. Learning-based MVS methods aggregate cost volume from different viewpoints and use neural networks for cost regularization, which achieve superior performance compared with traditional methods.

Recently, cascade-based architectures [7, 14, 54] have been widely applied. They compute different resolution depth maps in a coarse-to-fine manner and progressively narrow hypothesis plane guidance to reduce computational complexity. However, these approaches do not take advantage of valuable insight contained in early phases and only consider the pixel-wise depth attribute. Some methods, e.g. deformable kernel-based [47] and transformer-based [4, 8, 22, 27, 46], introduce finely designed external structures for feature extraction but do not fully exploit the geometric clues embedded in the MVS scenarios.

Unlike existing works, we propose to explore the geometric structures embedded in coarse stages for delicate estimations in finer stages. In particular, we build a two-branch fusion network to integrate geometric priors contained in coarse depth maps with ordinary features extracted by the classic FPN [23], and the fused geometry awareness features can provide solid foundations for robust aggregation. Meanwhile, coarse probability volumes with abundant geometric structures are embedded into the regularization network, and we replace the heavy 3D convolution with enhanced 2D regularization without degrading the quality of depth-wise correlation, resulting in lightweight but robust cost matching. However, MVS networks tend to produce severe misestimation at high-frequency clutter textures due to confused matching in coarse stages, which inevitably affects explicit geometry perception. We are inspired by the human behavior that a nearsighted person can still perceive a scene well without glasses, even if the texture details cannot be seen clearly. Based on the observation, we refer to the idea of curriculum learning [2] to embed coarse geometric priors into finer stages from easy to difficult. Specifically, we utilize the frequency domain filtering strategy to effectively alleviate redundant high-frequency textures without producing more learning parameters and leverage geometric structures embedded in different hierarchies of frequency for gradually delicate depth estimation.

In addition, depth ranges of MVS scenarios are often concentrated in several intervals, for this, we adopt the Gaussian-Mixture Model to simulate full-scene depth distribution and PauTa Criterion [31] allows us to depict loca-
tions that are too close or too far hidden in the long tailing of the depth distribution curve, e.g. sky. The depth distribution loss is proposed finally for full-scene similarity supervision.

In summary, the main contributions are as follows:

- We propose the geometric prior guided feature fusion and the probability volume geometry embedding approaches for robust cost matching.
- We enhance geometry awareness via the frequency domain filtering strategy and adopt the idea of curriculum learning for progressively introducing geometric clues from easy to difficult.
- We model the depth distribution of MVS scenarios using the Gaussian-Mixture Model assumption and build the full-scene geometry perception loss function.
- The proposed method is extensively evaluated on the DTU dataset and both intermediate and advanced sets of Tanks and Temples benchmark, all achieving brand-new state-of-the-art performance.

2. Related Works

Learning-based MVS Methods. Existing MVS methods can be classified into four categories: volumetric [20, 39], direct point cloud-based [11, 21], mesh-based [10], and depth map-based [3, 12, 36, 37, 50]. Among them, depth map-based methods decouple the complicated reconstruction task into per-view estimation and multi-view fusion, which have stronger flexibility. Recently, learning-based methods have shown remarkable progress over traditional methods. MVSNet [56] constructs the cost volume by aggregating deep features and camera parameters, and uses 3D CNN for regularization. And to reduce memory consumption, many follow-up works have been developed. R-MVSNet [57] adopts GRUs to regularize the cost volume in a sequential manner but leading to increased run-time. CasMVSNet [14], UCS-Net [7], and CVP-MVSNet [54] adopt cascade cost volumes or cost volume pyramid to estimate depth maps in a coarse-to-fine manner.

Improvements for MVS in Post-pyramid Era. Starting from [7, 14, 54], the improvement of learning MVS has entered the era of the pyramid model. Similar ideas are later explored to lower the GPU cost of 3D regularization or increase depth quality, such as coarse-to-fine depth optimization [27, 28, 32, 43, 45, 49, 51–53, 61], attention-based feature aggregation [25, 47, 55, 59, 60, 62, 63], and patch matching-based methods [13, 24, 44]. In addition, several other innovations have been applied to solve the MVS problem [22, 29, 48]. MVSNet++ [6] integrates the curriculum learning framework into the training process. TransMVS-Net [8], EPP-MVSNet [27], and MVSTER [46] either put forward a feature matching transformer to aggregate long-range context information or use the epipolar transformer to learn semantics and spatial associations. MVSFormer [4] proposes a pre-trained vision transformer to enhance the network. Although the popularity of the transformer [42] has inspired lots of downstream tasks, the fine-tuning vision transformer is sophisticated and does not fully explore the identities of the MVS problem. In this paper, we explore embedding geometric priors from coarse stages to analyze and exploit full-scene geometry awareness explicitly.

3. Methodology

Input a set of unstructured calibrated images, let \( I_0 \) be the reference image and \( \{ I_i \}_{i=1}^N \) as source images. GeoMVSNet estimates the depth map \( D \) with width \( W \) and height \( H \) alignment with \( I_0 \) through the collaboration of \( \{ I_i \}_{i=0}^{N} \).

The overall architecture of our network is illustrated in Fig. 1. Along the horizontal data flow, deep image features \( \{ F_i \}_{i=0}^N \) extracted from input images are first warped into the fronto-parallel planes of the reference camera frustum, denote as \( \{ V_i \}_{i=0}^N \). Then multiple feature volumes are aggregated into a single cost volume \( C \in \mathbb{R}^{G \times M \times H \times W} \), where \( G \) is the group-wise correlation channel [15]. Afterward, lightweight cost regularization is applied to \( C \) to obtain the probability volume \( P \in \mathbb{R}^{M \times H \times W} \), which represents the possibility of a pixel “sticking” to a depth plane.

Below, we first focus on the problem of robust cost matching and propose the geometric prior guided feature fusion and the probability volume geometry embedding in Sec. 3.1. We further extend the geometric clues in the frequency domain and continually enhance geometry perception through curriculum learning in Sec. 3.2, and finally describe the depth distribution similarity loss based on the Gaussian-Mixture Model in Sec. 3.3 and Sec. 3.4.

3.1. Geometry Awareness for Robust Cost Matching

The aggregation and regularization processes of cost matching in MVSNet [56] and related extension works [44, 55, 57] are much more robust than traditional MVS methods [3, 37, 50] which utilize normalized cross-correlation (NCC) to measure image patch similarity. However, in the most popular cascade MVS schemes [14, 54], image features and cost volumes of different stages often share the same constituents which do not fully explore the extensive geometric information supplied by early phases. Unlike existing works that rely on onerous external dependencies, we propose to explicitly fuse the geometric priors from coarse depth estimations and embed coarse probability volumes of coarser stages into cost matching at finer stages.

Geometric prior guided feature fusion. Take the \( \ell \) and \( \ell + 1 \) level as an example, the geometric prior guided feature of the reference image in the finer stage is formulated as

\[
Branch(z) = \hat{B}([D_{\ell+1}^f, B([I_0^{\ell+1}, D_{\ell+1}^f])]),
\]  

(1)
Figure 1. Illustration of GeoMVSNet. Structural features are extracted first by the geometry fusion network (Sec. 3.1) in finer stages, and W×A which denotes homography warping and aggregation is used to construct cascade cost volumes. The coarse probability volumes in coarse stages are embedded into the lightweight regularization network for geometry awareness (Sec. 3.1). And the frequency domain filtering equipped with curriculum learning strategy (Sec. 3.2) and depth distribution similarity loss (Sec. 3.4) based on Gaussian-Mixture Model (Sec. 3.3) are applied for full-scene geometry enhancement. The geometric prior output from the previous stage is used to guide the geometry perception for finer stages as shown by the numerical labels $0 \sim 2$.

$$F_{0}^{\ell+1}(z) = Fusion\{F_{0}^{\ell+1}(z) \oplus Branch(z)\},$$

where $z$ denotes image pixel, $[,]$ and $\oplus$ represent the concatenation and element-wise addition operation respectively. Two-branch network architecture is well studied in depth completion tasks [17, 38], and here we fuse the texture of the reference image and the upsampled geometric prior in the previous coarse depth by two neural submodules $\bar{B}$ and $\hat{B}$, and term the combination as Branch in Equ. 1. Then, the feature $F_{0}^{\ell+1}$ from classic FPN [23] is merged through the Fusion network. The architecture of the geometry fusion network for structural feature extraction is presented in Fig. 2. We can clearly see that the geometric prior aligned with the reference image is explicitly encoded into the basic FPN feature, and the geometric fused reference feature can be robustly matched with anisotropy source features.

Probability volume geometry embedding. As aforementioned, the probability volume $P$ represents the possibility that the depth of a certain pixel attaches to a depth hypothesis. Existing pyramid-based methods do not take advantage of a great deal of insight contained in $\{P_{i}\}_{i=0}^{N}$, but only use the coarse depth map it derived to reduce the computational consumption of denser space divisions. Since the scale and spatial extent of probability volumes vary from different stages, we use $\{P_{i}\}_{i=0}^{N}$ as the 3D “position maps” [9] embedding in the cost regularization network, without fragmenting them into the cost volume construction like what feature fusion does. In particular, we reduce the convolution kernel size from $k \times k \times k$ to $1 \times k \times k$ in the 3D cost regularization network, where the first dimension represents depth orientation. Meanwhile, the deficit caused by the lack of bulky but capable 3D convolutions in the depth direction is compensated by the explicitly coarse probability volume embedding. $P^{\ell}$ from the previous coarse stage is first passed through several 3D Maxpooling layers to construct the geometry perception pyramid with different sparsity rates. They are then explicitly encoded into different large receptive field hunting and skip connections layers of the U-Net [35] shape lightweight regularization network to build the fused spatial correlation. The 3D geometry position embedding can be mathematically expressed as

$$\begin{align*}
X &= \frac{(u - u_{0})Z}{f_{x}}, \\
Y &= \frac{(v - v_{0})Z}{f_{y}}, \\
Z &= \text{Prob}\{\{m\} \leftarrow M\},
\end{align*}$$

in which $(u, v)$ are the pixel coordinates of a voxel in $m$-th candidate hypothesis from $M$ pre-defined total depth planes, and $u_{0}, v_{0}, f_{x}, f_{y}$ are part of camera intrinsic parameters. The geometry embedding of probability volumes...
from coarse-to-fine stages is shown in Fig. 3, where geometry embedding slices along the depth direction encode spatial perception about the overall structure of the scene, rather than just providing pixel-level “confidence”. Fine-grained geometry awareness is continuously passed through the network for robust cost matching.

Geometric prior guided feature fusion strengthens the discrimination and structure of deep features at finer stages without introducing external complex dependencies, laying a solid foundation for robust aggregation. Embedded probability volumes not only provide voxel coordinates and depth-aware positional encoding for robust cost volume regularization but also introduce full-scene depth distribution characteristics into the depth perception of finer layers.

3.2. Geometry Enhancement in Frequency Domain

Coarse depth map fusion and probability volume embedding can effectively integrate progressively enhanced geometry awareness into cost matching. Despite of this, severe misestimations at clutter textures inherent in the coarse depth map, e.g. infinite sky in the frame and areas near the edge of the image where reprojection is extremely prone to out of bounds, inevitably increase the learning burden of the Fusion network and cost regularization network.

To fix severely erroneous depth values, we attempt to use the pre-trained RGB-guided depth refinement modules [64]. The plug-in depth optimization module can indeed polish the depth map visually, especially at the object contours. However, performing a spectral analysis of the depth map in

![Figure 2. The architecture of the geometry fusion network.](image)

Figure 2. The architecture of the geometry fusion network. The coarse depth is used as the geometric prior of two branches. Specific data structures and parameters can be found in Supp. II.1.

However, performing a spectral analysis of the depth map in the depth map visually, especially at the object contours. The plug-in depth optimization module can indeed polish the pre-trained RGB-guided depth refinement modules [64].

In contrast, we approach the problem using frequency domain filtering [33] via the Discrete Fourier Transform (DFT) [40]. We regard the coarse depth map as a 2D discrete signal and transform it to the frequency domain by Equ. 4, where $j$ is the imaginary unit.

$$\mathcal{F}^j(u, v) = \sum_{x=0}^{W-1} \sum_{y=0}^{H-1} D^j(x, y) e^{-j2\pi(xu + yv)}. \quad (4)$$

$$\tilde{D}^j(x, y) = \frac{1}{WH} \sum_{u=0}^{W-1} \sum_{v=0}^{H-1} \mathcal{F}^j(u, v) e^{j2\pi(xu + yv)}. \quad (5)$$

After FFT-shift [1], a basic ideal rectangular low-pass filter [41] is used to eliminate high-frequency information from the coarse depth map as shown in Fig. 4 (b), and Equ. 5 is used for inverse domain transfer. The simple but effective frequency filtering ingeniously removes the complex and incomprehensible knowledge from the explicitly modeled coarse geometry embedding while avoiding producing more learning parameters. Meanwhile, severe misestimation and high-frequency burden signals are alleviated without using hand-labeled visual masks, allowing the network to focus more on the full-scene geometry perception.

We also refer to the idea of curriculum learning [2], incrementally teaching difficult depth embedding samples to the Fusion network and cost regularization network. Let $d^j$ define the random variable of estimated depth map $D^j$ at

![Figure 3. Visualization of the probability volume geometry embedding on the Tanks and Temples dataset [19].](image)

Fig. 4 (a), we find that the polished depth has significantly higher frequency information, which burdens the network learning [26,51]. More importantly, the seemingly accurate “refinement” operation reduces the satisfaction of geometric consistency constraints, leading to significantly deteriorated overall quality for the point cloud (0.200 v.s. 0.704). The main reason is that RGB-guided depth optimization tends to fit depth distributions in the dataset, while MVS estimates geometrically consistent depths by matching.

After FFT-shift [1], a basic ideal rectangular low-pass filter [41] is used to eliminate high-frequency information from the coarse depth map as shown in Fig. 4 (b), and Equ. 5 is used for inverse domain transfer. The simple but effective frequency filtering ingeniously removes the complex and incomprehensible knowledge from the explicitly modeled coarse geometry embedding while avoiding producing more learning parameters. Meanwhile, severe misestimation and high-frequency burden signals are alleviated without using hand-labeled visual masks, allowing the network to focus more on the full-scene geometry perception.

We also refer to the idea of curriculum learning [2], incrementally teaching difficult depth embedding samples to the Fusion network and cost regularization network. Let $d^j$ define the random variable of estimated depth map $D^j$ at
the \( \ell \)-th stage, and the target distribution of the scenario is \( \mathcal{N} \). Let \( 0 \leq W^\ell(d^\ell) \leq 1 \) be the weight applied to example \( d^\ell \) in the curriculum sequence. The training distribution is

\[
Q^\ell(d^\ell) \propto W^\ell(d^\ell) \mathcal{N}(d^\ell), \; d^\ell \in D^\ell.
\]  

(6)

We adjust the monotonically increasing weight \( W^\ell \) by modulating the cutout kernel ratio of frequency domain filter, denoted as \( \rho \), and leave the geometric clues untrimmed (\( \rho = 1 \)) at the last stage of the coarse-to-fine scheme [14,54]. The curriculum learning strategy as shown in Fig. 4 (c) introduces a better geometric clues consumption pattern for the cost regularization network, effectively enhancing the full-scene geometry awareness for the MVS network.

### 3.3. Mixed-Gaussian Depth Distribution Model

Given a pre-estimated depth range \([d_{\text{min}}, d_{\text{max}}]\) from sparse reconstruction by classic structure-from-motion algorithms [5, 36], existing learning-based MVS methods [14, 56] always follow the uniform depth distribution assumption that divides the reference camera frustum into \( M \) depth hypothesis planes. CIDER [51] proposes to partition the hypothesis planes in the inverse depth space, and Yang et al. [53] introduce the multi-modal depth distribution. However, these methods only consider pixel-wise depth characteristics and do not model the full-scene depth distribution, which is pivotal for geometry perception.

The scenarios to be reconstructed in current studies can be divided into three categories: a) centered object and orbiting camera; b) surrounding object and self-rotating camera; c) aerial photograph. Fig. 5 visualizes the image and depth distribution for each category. The depth range of natural scenes is often concentrated in several certain areas, and locations that are too close and too far are hidden in the long trailing of the depth distribution curve.

Based on the observation, we assume the random variable depth value \( d \) follows the Gaussian-Mixture Model (GMM) distribution [34]. The sample distribution can be modeled as \( \mathcal{N}(d; \mu_i, \sigma_i^2) \), where \( \theta_i = \{\mu_i, \sigma_i\} \) are the mean and standard deviation of the \( i \)-th Gaussian component respectively. And the probability density function is given by

![Figure 5. Full-scene depth distribution of scenarios in three categories on the BlendedMVS dataset [58]. Most scenarios can be modeled by the GMM with \( K \leq 2 \), and the details about distribution histograms and fitting curves can be found in Supp. I.](image-url)


\[ p(d \mid \Omega) = \sum_{i=1}^{K} \omega_i \Phi(d \mid \theta_i), \quad (7) \]

\[ \Phi(d \mid \theta_i) = \frac{1}{\sqrt{2\pi\sigma_i}} \exp\left(-\frac{(d - \mu_i)^2}{2\sigma_i^2}\right), \quad (8) \]

where \( \Omega = \{\omega_i\}, i = (1, 2, ..., K) \) is the set of prior distributions modeling the probability that variable \( d \) falls in the approximate estimation interval, satisfying the constraints

\[ 0 \leq \omega_i \leq 1 \quad \text{and} \quad \sum_{i=1}^{K} \omega_i = 1. \quad (9) \]

We find that most scenarios are well portrayed at \( K = 1 \) or 2, and only a few scenarios are modeled by the distribution at \( K \geq 3 \) (see Supp. I). And the PauTa Criterion [31] allows us to depict the depth distribution of the whole scene well within the combination of several \( (\mu_i - 3\sigma_i, \mu_i + 3\sigma_i) \) intervals. The long-standing burden of infinite points (e.g., sky) will not bring a negative impact on learning under the GMM with PauTa Criterion, and we can better use the full-scene depth distribution to enhance spatial perception.

### 3.4. Loss Functions

Pixel-wise classification modeling [47, 52] is more suitable for our representation since regression mode [14, 56] and recently published unification mode [32] tend to fall among \( 2 \times 2 \) analysis. The number of hypothesis planes \( M \) and the number of hypothesis planes \( M = 4 \) for each level respectively. The depth sampling range is \( 425 \text{mm} \sim 935 \text{mm} \) for DTU and \( (\mu_i, \sigma_i) \) are self-calculated according to each scene. The cutout filter kernel ratio \( \rho \) is set to \( 9, 4, 2, 1 \) as shown in Fig. 4 (c), and the weight allocation for loss items is mentioned in Eq. 13. We use PyTorch [30] for implementation and train the model with the Adam optimizer for 16 epochs from a start learning rate of 0.001 on 2 NVIDIA Tesla V100 GPUs.

### 4. Experiments

#### 4.1. Datasets

**DTU** [18] is an indoor dataset consisting of 124 different objects, each scene is recorded from 49 views with 7 brightness levels. It contains ground-truth point clouds collected under well-controlled laboratory conditions for evaluation.

**Tanks and Temples (T&T)** [19] dataset contains a more challenging realistic environment with large-scale variations and illumination changes. It contains an intermediate subset of 8 scenes and an advanced subset of 6.

**BlendedMVS** [58] dataset is a recently published large-scale synthetic dataset. It consists of over 17000 high-resolution rendered images with 3D structures.

#### 4.2. Implementation Details

Following the common practice, we train the GeoMVS-Net on the DTU [18] training set and evaluate it on the DTU evaluation set while adopting the same data split and view selection as defined in [56] and [14] for a fair comparison. And we train our model on the BlendedMVS dataset [58] and test on both intermediate and advanced sets of the Tanks and Temples benchmark [19].

**Training.** The number of input images is set to \( N = 5 \) with a resolution of \( 640 \times 512 \) for the DTU, and \( N = 7 \) with \( 768 \times 576 \) images for the BlendedMVS. We use \( L = 4 \) layer pyramids and the number of hypothesis planes \( M = 4 \) for each level respectively. The depth sampling range is \( 425 \text{mm} \sim 935 \text{mm} \) for DTU and \( (\mu_i, \sigma_i) \) are self-calculated according to each scene. The cutout filter kernel ratio \( \rho \) is set to \( 9, 4, 2, 1 \) as shown in Fig. 4 (c), and the weight allocation for loss items is mentioned in Eq. 13. We use PyTorch [30] for implementation and train the model with the Adam optimizer for 16 epochs from a start learning rate of 0.001 on 2 NVIDIA Tesla V100 GPUs.

**Evaluation.** Other settings are consistent with the training process, except for input image properties. We crop the image to \( 1024 \times 1024 \) and also use \( N = 5 \) for the DTU evaluation. We resize the height of the T&T images to 1024 while remaining the width to 1920 or 2048 unchanged according to different testing scenes and use \( N = 11 \) input views. Our model consumes 0.26s and 5.98G memory for the full-resolution DTU depth estimation and 0.47s and 8.85G memory for the T&T. As for depth fusion, we use the open-source 3D data processing library Open3D [65] for dense point cloud fusion for the DTU, and adopt the commonly used dynamic fusion strategy [52] for the T&T. It is worth noting that we do not elaborately tune the fusion parameters, but fuse the full-scene point cloud using pixels with confidence \( c \geq \mu - 3\sigma \) for each scenario on the assumption of the GMM at \( K = 1 \) in Sec. 3.3.

**Metrics.** For point cloud evaluation, the accuracy and completeness of the distance metric are adopted for DTU [18]
while the accuracy and completeness of the percentage metric for T&T [19]. Besides, there is an official website for online evaluation of Tanks and Temples [19] benchmark.

4.3. Benchmark Performance

DTU. We compare our results with traditional methods and recent learning-based methods. The qualitative results are shown in Fig. 6. GeoMVSNet estimates significantly accurate depths and complete point clouds, especially for the geometry structures of the subject, and high-frequency clutter textures are well suppressed. Meanwhile, scan48 and scan77 with drastic illumination changes and reflections are considered as two most difficult scenes on the DTU evaluation set, which further proves the robustness of our method.

For quantitative evaluation, we report accuracy and completeness using official MATLAB codes as shown in Tab. 1. Our approach outperforms all current methods in completeness and raises the overall metrics to a new altitude.

Tanks and Temples. We further validate the generalization ability of our proposed method on the T&T dataset. Fig. 7 shows the error comparison of the reconstructed point clouds, our method has higher precision and recall, especially in geometrically informative regions. And the quantitative results on both intermediate and advanced sets are reported in Tab. 2, our method achieves state-of-the-art performance among all existing MVS methods and yields first place in most scenes. In particular, we rank first among all submissions on the advanced set, demonstrating our robustness and generalization performance on large and challenging MVS scenarios. Visualization of more reconstructed point clouds can be found in Supp. III.2.

4.4. Ablation Study

Tab. 3 shows the ablation results of our proposed GeoMVSNet. The baseline [14] method is re-customized according to the number of pyramid layers and input view numbers. However, the geometric clues embedded in coarse stages are not exploited.

Effect of geometry awareness. The geometry fusion network which utilizes the geometric prior derived from the

### Table 1. Quantitative comparison on the DTU dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>Acc. (mm)</th>
<th>Comp. (mm)</th>
<th>Overall↓ (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gipuma [12]</td>
<td>0.283</td>
<td>0.873</td>
<td>0.578</td>
</tr>
<tr>
<td>COLMAP [36]</td>
<td>0.400</td>
<td>0.664</td>
<td>0.532</td>
</tr>
<tr>
<td>R-MVSNet [57]</td>
<td>0.383</td>
<td>0.452</td>
<td>0.417</td>
</tr>
<tr>
<td>CasMVSNet [14]</td>
<td>0.325</td>
<td>0.385</td>
<td>0.355</td>
</tr>
<tr>
<td>CVP-MVSNet [54]</td>
<td>0.296</td>
<td>0.406</td>
<td>0.351</td>
</tr>
<tr>
<td>EPP-MVSNet [27]</td>
<td>0.413</td>
<td>0.296</td>
<td>0.355</td>
</tr>
<tr>
<td>CER-MVS [28]</td>
<td>0.359</td>
<td>0.305</td>
<td>0.332</td>
</tr>
<tr>
<td>RayMVSNet [48]</td>
<td>0.341</td>
<td>0.319</td>
<td>0.330</td>
</tr>
<tr>
<td>Eff-MVSNet [45]</td>
<td>0.321</td>
<td>0.313</td>
<td>0.317</td>
</tr>
<tr>
<td>CDS-MVSNet [13]</td>
<td>0.352</td>
<td>0.280</td>
<td>0.316</td>
</tr>
<tr>
<td>NP-CVP-MVSNet [53]</td>
<td>0.356</td>
<td>0.278</td>
<td>0.315</td>
</tr>
<tr>
<td>UniMVSNet [32]</td>
<td>0.352</td>
<td>0.278</td>
<td>0.315</td>
</tr>
<tr>
<td>TransMVSNet [8]</td>
<td>0.321</td>
<td>0.289</td>
<td>0.305</td>
</tr>
<tr>
<td>GBi-Net* [29]</td>
<td>0.312</td>
<td>0.293</td>
<td>0.303</td>
</tr>
<tr>
<td>MVSTER* [46]</td>
<td>0.340</td>
<td>0.266</td>
<td>0.303</td>
</tr>
<tr>
<td>GeoMVSNet (Ours)</td>
<td>0.331</td>
<td>0.259</td>
<td>0.295</td>
</tr>
</tbody>
</table>
Table 2. Quantitative results on the Tanks and Temples dataset. Bold represents the best while underlined represents the second-best.

<table>
<thead>
<tr>
<th>Method</th>
<th>Intermediate</th>
<th>Advanced</th>
</tr>
</thead>
<tbody>
<tr>
<td>COLMAP [36]</td>
<td>42.14</td>
<td>50.11</td>
</tr>
<tr>
<td>CasMVSSNet [14]</td>
<td>56.42</td>
<td>76.36</td>
</tr>
<tr>
<td>PatchMatchNet [44]</td>
<td>53.15</td>
<td>66.99</td>
</tr>
<tr>
<td>CER-MVS [28]</td>
<td>64.82</td>
<td>81.16</td>
</tr>
<tr>
<td>EfS-MVSSNet [45]</td>
<td>56.88</td>
<td>72.21</td>
</tr>
<tr>
<td>UmMVSNet [32]</td>
<td>64.36</td>
<td>81.20</td>
</tr>
<tr>
<td>TransMVSNet [8]</td>
<td>63.52</td>
<td>80.92</td>
</tr>
<tr>
<td>GBi-Net [29]</td>
<td>61.42</td>
<td>79.77</td>
</tr>
<tr>
<td>MVSTER [46]</td>
<td>60.92</td>
<td>80.21</td>
</tr>
<tr>
<td>GeoMVSNet (Ours)</td>
<td>65.89</td>
<td>81.64</td>
</tr>
</tbody>
</table>

Figure 7. Point clouds error comparison of state-of-the-art methods on the Tanks and Temples benchmark. τ is the scene-relevant distance threshold determined officially and darker means larger error. The first row shows Precision and the second row shows Recall.

Table 3. Ablation results on the DTU evaluation dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>Sec.1.1</th>
<th>Sec.1.2</th>
<th>Sec.1.4</th>
<th>Acc.</th>
<th>Comp.</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (L=4, N=5)</td>
<td>GPN</td>
<td>PVE</td>
<td>GFN</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ geometry fusion network</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>0.9629</td>
<td>0.9016</td>
<td>0.9325</td>
</tr>
<tr>
<td>+ prob. volume embedding</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>0.9352</td>
<td>0.2493</td>
<td>0.9207</td>
</tr>
<tr>
<td>+ fusion &amp; embedding</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>0.3705</td>
<td>0.3053</td>
<td>0.3739</td>
</tr>
<tr>
<td>+ frequency domain filtering</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>0.3404</td>
<td>0.2922</td>
<td>0.3683</td>
</tr>
<tr>
<td>+ curriculum learning</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>0.3663</td>
<td>0.2307</td>
<td>0.3185</td>
</tr>
<tr>
<td>+ distribution similarity loss</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>0.3550</td>
<td>0.2434</td>
<td>0.3142</td>
</tr>
<tr>
<td>proposed</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>0.3392</td>
<td>0.2683</td>
<td>0.2989</td>
</tr>
</tbody>
</table>

5. Conclusion

In this paper, we propose GeoMVSNet which explicitly integrates coarse geometry structures into finer depth estimations, achieving prominent geometry perception for MVS scenarios. Specifically, we construct a two-branch feature fusion network to fuse geometric priors from coarse stages with basic unstructured features and embed coarse probability volumes into the lightweight cost regularization network for geometry awareness without introducing complicated external dependencies. In addition, we utilize frequency domain filtering to suppress high-frequency clutter misestimations and the curriculum learning strategy further introduces a better geometric information consumption pattern for robust cost matching. And the proposed depth distribution similarity loss based on the Gaussian-Mixture Model assumption enhances the full-scene depth perception. We achieve state-of-the-art performance on both DTU and Tanks and Temples datasets and rank first on the T&T-Advanced set. In the future, we intend to discover the ability of explicitly modeled geometry extensions in the field of unsupervised or self-supervised MVS frameworks.

Acknowledgements. This research is supported by the National Natural Science Foundation of China U21B2012 and 62072013, Shenzhen Cultivation of Excellent Scientific and Technological Innovation Talents RJC202000714114435057, Shenzhen Science and Technology Program - Shenzhen Hong Kong joint funding project of SGD2021123144400001, and the Outstanding Talents Training Fund in Shenzhen.
References


