GeoMVSNet: Learning Multi-View Stereo with Geometry Perception

Zhe Zhang[1] Rui Peng[1] Yuxi Hu[2] Ronggang Wang[1]
[1]School of Electronic and Computer Engineering, Peking University, China
[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 locations 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. Cas-MVSNet [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. TransMVSNet [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 ℓ and $\ell+1$ level as an example, the geometric prior guided feature of the reference image in the finer stage is formulated as

$$Branch(z) = \hat{\mathcal{B}}([D_{\uparrow}^{\ell}, \mathcal{B}([I_0^{\ell+1}, D_{\uparrow}^{\ell}])]) \, , \tag{1}$$

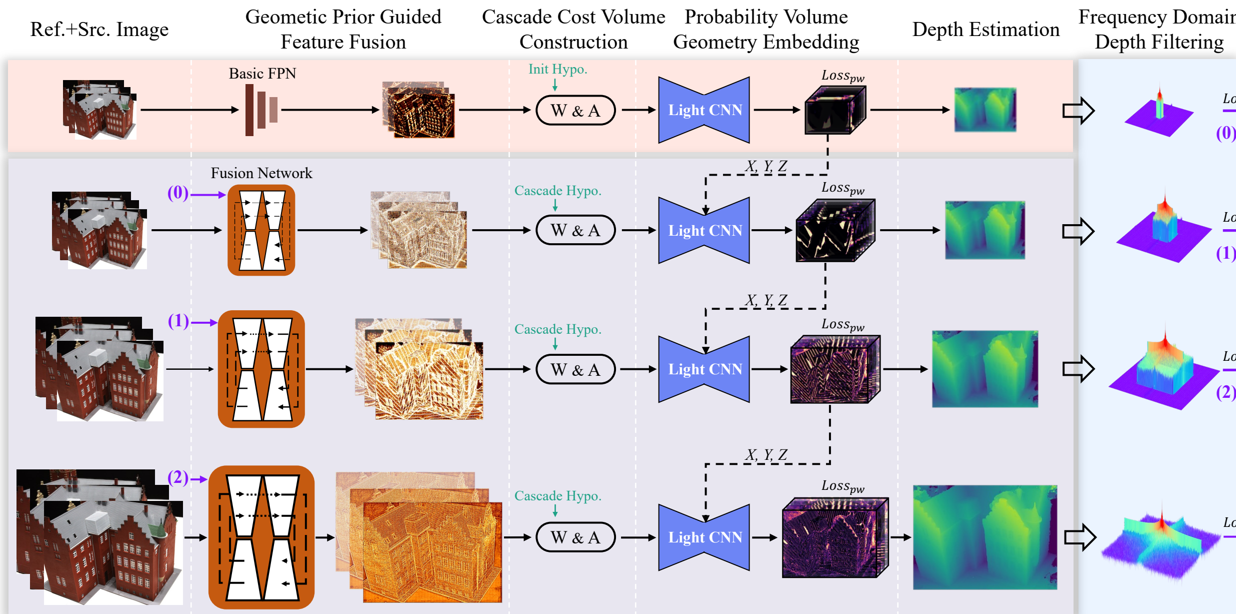


Figure 1. **Illustration of GeoMVSNet.** Structural features are extracted first by the geometry fusion network (Sec. 3.1) in finer stages, and W&A which denotes homography warping and aggregation is used to construct cascade cost volumes. The coarse probability volumes in coarse stages are embedded 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).

$$F_0^{\ell+1}(z) = Fusion\{\bar{F}_0^{\ell+1}(z) \oplus Branch(z)\}\ , \qquad (2)$$

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 $\mathcal{B}$ and $\hat{\mathcal{B}}$, and term the combination as *Branch* in Equ. 1. Then, the feature $\bar{F}_0^{\ell+1}$ 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^ℓ 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

$$\begin{cases} X = \dfrac{(u-u_0)Z}{f_x} \\ Y = \dfrac{(v-v_0)Z}{f_y} \\ Z = Prob(\{m\} \leftarrow M) \end{cases}\ , \qquad (3)$$

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

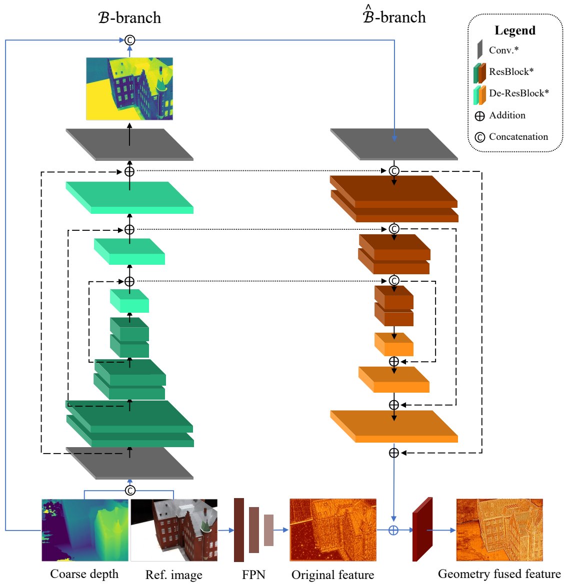


Figure 2. **The architecture of the geometry fusion network.** The coarse depth is used as the geometric prior of two branches. Specific data structures and parameters can be found in *Supp. II.1*.

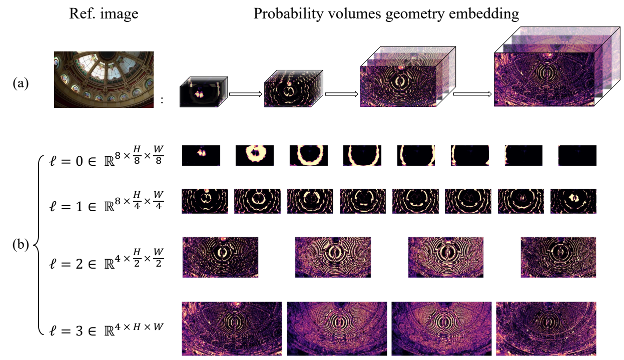


Figure 3. **Visualization of the probability volume geometry embedding on the Tanks and Temples dataset [19].** (a) Summary of the coarse-to-fine embedding process; (b) basics of geometry awareness derived from probability volumes of different stages. More examples are presented in *Supp. II.3*.

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 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.

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.

$$\mathcal{F}^{\ell}(u,v) = \sum_{x=0}^{W-1}\sum_{y=0}^{H-1} D^{\ell}(x,y)\, e^{-j2\pi\left(\frac{ux}{W}+\frac{vy}{H}\right)}\,. \tag{4}$$

$$\tilde{D}^{\ell}(x,y) = \frac{1}{WH}\sum_{u=0}^{W-1}\sum_{v=0}^{H-1} \tilde{\mathcal{F}}^{\ell}(u,v)\, e^{j2\pi\left(\frac{ux}{W}+\frac{vy}{H}\right)}\,. \tag{5}$$

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^{ℓ} define the random variable of estimated depth map D^{ℓ} at

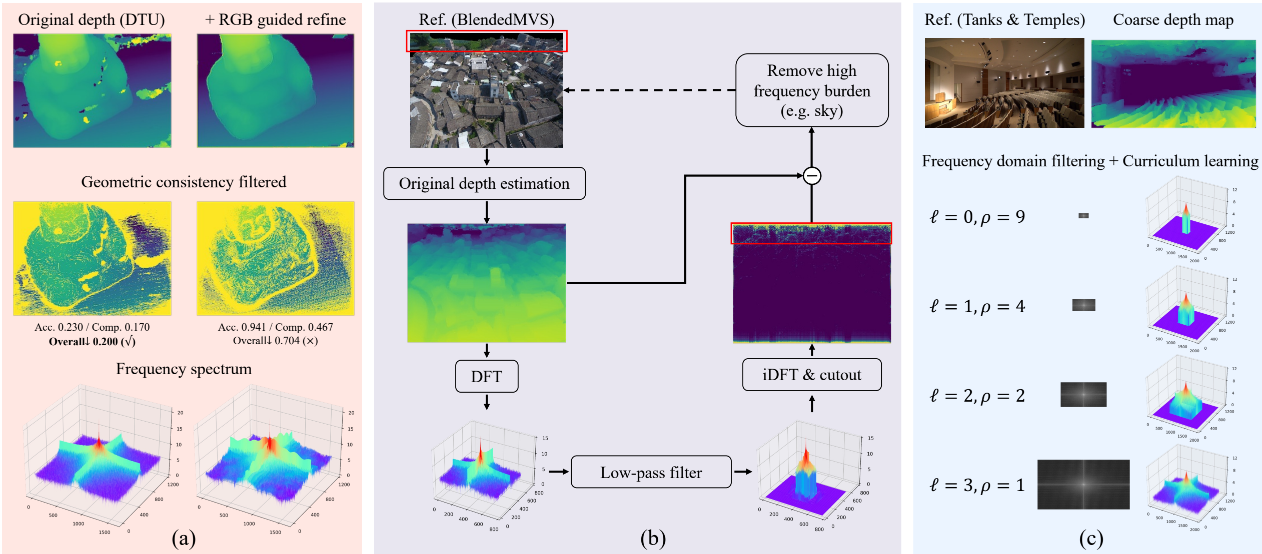


Figure 4. **Analysis of geometry enhancement in the frequency domain.** (a) The experiment results of the depth map refinement module on the DTU dataset [18]; (b) schematic flow chart of the frequency domain filtering on the BlendedMVS dataset [58]; (c) curriculum learning parameter configuration on the advanced set of Tanks and Temples dataset [19], coordinate spaces are unified for visualization.

the ℓ-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^ℓ in the curriculum sequence. The training distribution is

$$Q^\ell(d^\ell) \propto W^\ell(d^\ell)\,\mathcal{N}(d^\ell)\;,\; d^\ell \in D^\ell\;. \tag{6}$$

We adjust the monotonically increasing weight W^ℓ by modulating the cutout kernel ratio of frequency domain filter, denoted as ρ, 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 $[d_{min}, d_{max}]$ 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

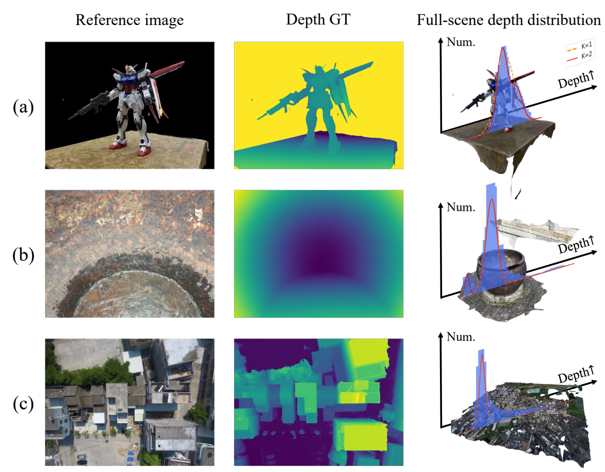


Figure 5. **Full-scene depth distribution of scenarios in three categories on the BlendedMVS dataset [58].** Most scenarios can be modeled by the GMM with $K \leq 2$, and the details about distribution histograms and fitting curves can be found in *Supp. I*.

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

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

$$\Phi(d \mid \theta_i) = \frac{1}{\sqrt{2\pi}\sigma_i} exp(-\frac{(d-\mu_i)^2}{2\sigma_i^2}) \, , \tag{8}$$

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

$$0 \leq \omega_i \leq 1 \;\; and \;\; \sum_{i=1}^{K} \omega_i = 1 \, . \tag{9}$$

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 into a local optimal solution at early stages. The cross-entropy loss for pixel-wise supervised is

$$Loss_{pw} = \sum_{z \in \Psi} (-P_{GT}(z) \; log[P(z)]) \, , \tag{10}$$

where Ψ denotes the set of valid pixels with ground truth precision, and P_{GT} is the ground-truth probability volume.

Moreover, the perception of the full-scene depth distribution is the focus of this paper. We calculate the similarity of sample distribution between filtered depth estimation and ground-truth depth using the Kullback-Leibler Divergence [16] metric. The depth distribution similarity loss is

$$Loss_{dds} = \sum_{m=0, z \in \Upsilon}^{M'} \tilde{p}(z) \; (log \; \tilde{p}(z) - log \; \mathcal{N}_{GT}(z)) \, , \tag{11}$$

$$\Upsilon = \Psi \, \cap \, \bigcup_{i=1}^{K} \{(\mu_i - 3\sigma_i, \mu_i + 3\sigma_i)\} \, , \tag{12}$$

where $\tilde{p}(:)$ denotes the filtered depth distribution, and we slice the depth space of each scene into $M' = 48$ discrete intervals to calculate the similarity of the depth distribution.

The overall loss is a weighted sum of $Loss_{pw}$ and $Loss_{dds}$ in Equ. 13, where $\lambda_1^* = 0.8$, $\lambda_2^* = 0.2$ among each stage in our experiments.

$$Loss = \sum_{\ell=0}^{L} (\lambda_1^{\ell} \; Loss_{pw} + \lambda_2^{\ell} \; Loss_{dds}) \, . \tag{13}$$

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 GeoMVSNet 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 640×512 for the DTU, and $N = 7$ with 768×576 images for the BlendedMVS. We use $L = 4$ layer pyramids and the number of hypothesis planes M is set to $8, 8, 4, 4$ for each level respectively. The depth sampling range is $425mm \sim 935mm$ for DTU and (μ_i, σ_i) are self-calculated according to each scene. The cutout filter kernel ratio ρ is set to $9, 4, 2, 1$ as shown in Fig. 4 (c), and the weight allocation for loss items is mentioned in Equ. 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 1600×1152 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]

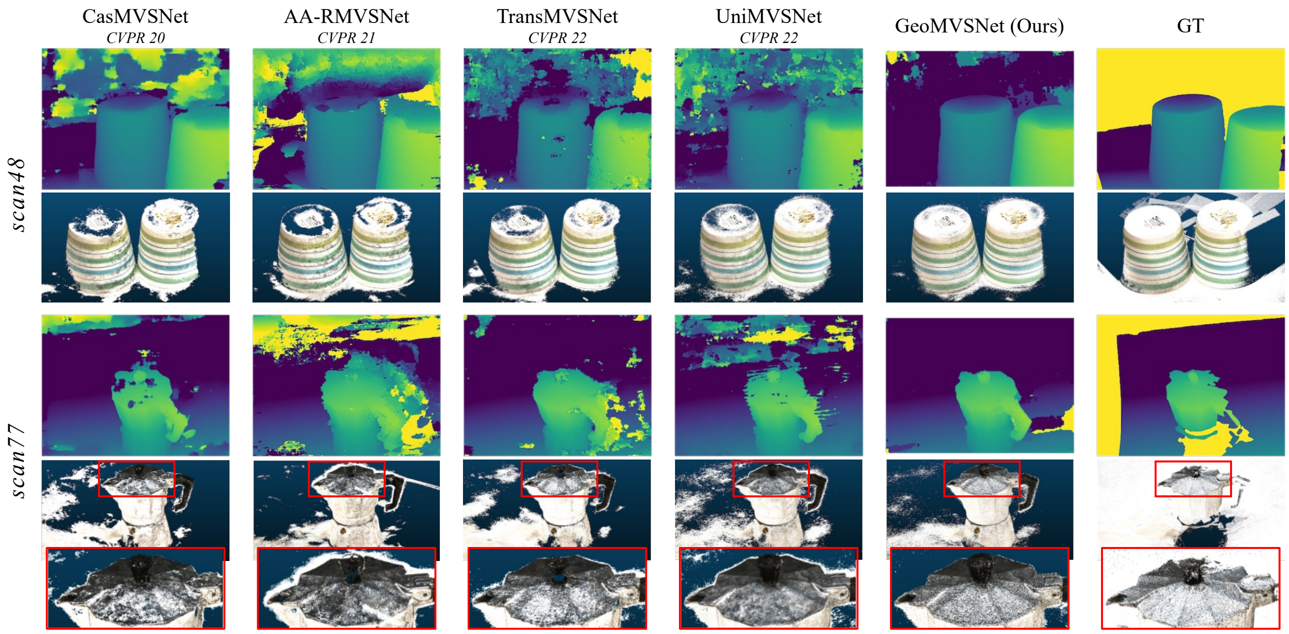


Figure 6. **Qualitative comparison of the most challenging *scan48* and *scan77* on the DTU evaluation dataset.** The first and third rows are estimated depth maps while others are point cloud reconstruction results. Our model produces remarkable accuracy and completeness.

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*.

Table 1. **Quantitative comparison on the DTU dataset.** $\star$ means that GBi-Net [29] is re-tested with the same post-processing threshold on all scans for fair comparisons with other methods. $*$ means MVSTER [46] is trained on full-resolution images.

	Method	Acc. (*mm*)	Comp. (*mm*)	Overall↓ (*mm*)
	Gipuma [12]	**0.283**	0.873	0.578
	COLMAP [36]	0.400	0.664	0.532
	R-MVSNet [57]	0.383	0.452	0.417
	CasMVSNet [14]	0.325	0.385	0.355
	CVP-MVSNet [54]	0.296	0.406	0.351
Post-pyramid Era	EPP-MVSNet [27]	0.413	0.296	0.355
	CER-MVS [28]	0.359	0.305	0.332
	RayMVSNet [48]	0.341	0.319	0.330
	Effi-MVSNet [45]	0.321	0.313	0.317
	CDS-MVSNet [13]	0.352	0.280	0.316
	NP-CVP-MVSNet [53]	0.356	0.275	0.315
	UniMVSNet [32]	0.352	0.278	0.315
	TransMVSNet [8]	0.321	0.289	0.305
	GBi-Net⋆ [29]	0.312	0.293	0.303
	MVSTER* [46]	0.340	0.266	0.303
	GeoMVSNet (Ours)	0.331	**0.259**	**0.295**

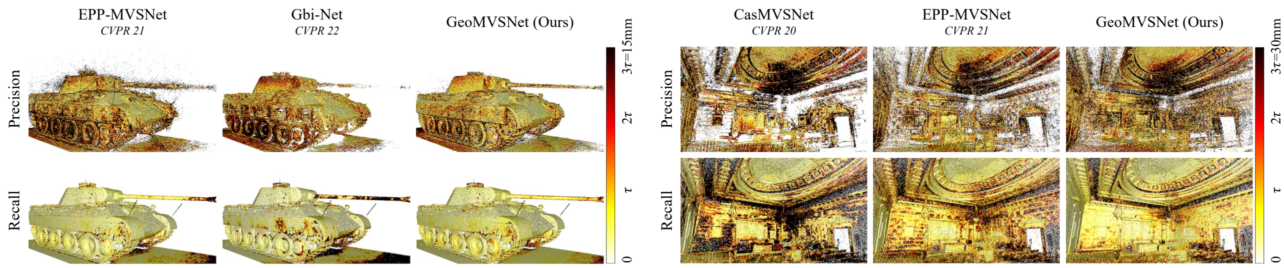
4.4. Ablation Study

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 2. **Quantitative results on the Tanks and Temples dataset. Bold** represents the best while <u>underlined</u> represents the second-best.

Method	**Intermediate**									**Advanced**						
	Mean↑	Family	Francis	Horse	L.H.	M60	Panther	P.G.	Train	**Mean↑**	Aud.	Bal.	Cou.	Mus.	Pal.	Tem.
COLMAP [36]	42.14	50.41	22.25	25.63	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 [14]	56.42	76.36	58.45	46.20	55.53	56.11	54.02	58.17	46.56	31.12	19.81	38.46	29.10	43.87	27.36	28.11
PatchmatchNet [44]	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
CER-MVS [28]	<u>64.82</u>	81.16	64.21	50.43	**70.73**	63.85	63.99	**65.90**	58.25	<u>40.19</u>	25.95	<u>45.75</u>	39.65	51.75	<u>35.08</u>	<u>42.97</u>
Effi-MVSNet [45]	56.88	72.21	51.02	51.78	58.63	58.71	56.21	57.07	49.38	34.39	20.22	42.39	33.73	45.08	29.81	35.09
UniMVSNet [32]	64.36	<u>81.20</u>	66.43	53.11	63.46	**66.09**	<u>64.84</u>	62.23	57.53	38.96	28.33	44.36	<u>39.74</u>	<u>52.89</u>	33.80	34.63
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
GBi-Net [29]	61.42	79.77	**67.69**	51.81	61.25	60.37	55.87	60.67	53.89	37.32	<u>29.77</u>	42.12	36.30	47.69	31.11	36.93
MVSTER [46]	60.92	80.21	63.51	52.30	61.38	61.47	58.16	58.98	51.38	37.53	26.68	42.14	35.65	49.37	32.16	39.19
GeoMVSNet (Ours)	**65.89**	**81.64**	<u>67.53</u>	<u>55.78</u>	<u>68.02</u>	<u>65.49</u>	**67.19**	<u>63.27</u>	<u>58.22</u>	**41.52**	**30.23**	**46.53**	**39.98**	**53.05**	**35.98**	**43.34**



Panther (Intermediate) *Courtroom* (Advanced)

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.**

Method	**Sec. 3.1**		**Sec. 3.2**		**Sec. 3.4**		**Acc.**	**Comp.**	**Overall↓**
	GFN	PVE	FDF	CL	$Loss_{pw}$	$Loss_{dds}$			
baseline (L=4, N=5)					✓		0.3629	0.3016	0.3323
+ geometry fusion network	✓				✓		0.3520	0.2893	0.3207
+ prob. volume embedding		✓			✓		0.3705	0.3053	0.3379
+ fusion & embedding	✓	✓			✓		0.3404	0.2922	0.3163
+ frequency domain filtering	✓		✓		✓		0.3663	0.2707	0.3185
+ curriculum learning	✓		✓	✓	✓		0.3650	0.2634	0.3142
+ distribution similarity loss	✓	✓			✓	✓	0.3346	0.2832	0.3089
proposed	✓	✓	✓	✓	✓	✓	**0.3309**	**0.2593**	**0.2951**

coarse depth map can significantly improve reconstruction completeness. However, embedding coarse probability volumes as 3D positional maps alone is insufficient to boost performance significantly. The proposed probability volume embedding strategy requires structural features as the foundation to achieve the best reconstruction quality.

Effect of frequency domain geometry enhancement. As shown in Tab. 3 and Fig. 4 (b), the frequency domain filtering approach equipped with the curriculum learning strategy can effectively eliminate clutter textures in coarse stages, preventing contamination from misestimated embedding and highlighting the effect of geometry awareness.

The full-scene depth distribution perception. Finally, the depth distribution similarity loss based on the assumption of the GMM and the combination of explicitly modeled geometric integration bring about an improvement in accuracy and fully depict the full-scene geometry perception.

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 RCJC20200714114435057, Shenzhen Science and Technology Program - Shenzhen Hong Kong joint funding project of SGDX20211123144400001, and the Outstanding Talents Training Fund in Shenzhen.

References

[1] Marwan Abdellah, Salah Saleh, Ayman Eldeib, and Amr Shaarawi. High performance multi-dimensional (2d/3d) fft-shift implementation on graphics processing units (gpus). In *2012 Cairo International Biomedical Engineering Conference (CIBEC)*, pages 171–174. IEEE, 2012.

[2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

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

[4] 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.

[5] Dan Cernea. OpenMVS: Multi-view stereo reconstruction library. 2020.

[6] Po-Heng Chen, Hsiao-Chien Yang, Kuan-Wen Chen, and Yong-Sheng Chen. Mvsnet++: Learning depth-based attention pyramid features for multi-view stereo. *IEEE Transactions on Image Processing*, 29:7261–7273, 2020.

[7] 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.

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

[9] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 534–551, 2018.

[10] Pascal Fua and Yvan G Leclerc. Object-centered surface reconstruction: Combining multi-image stereo and shading. *International Journal of Computer Vision*, 16(1):35–56, 1995.

[11] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2009.

[12] 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.

[13] Khang Truong Giang, Soohwan Song, and Sungho Jo. Curvature-guided dynamic scale networks for multi-view stereo. *arXiv preprint arXiv:2112.05999*, 2021.

[14] 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.

[15] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3273–3282, 2019.

[16] John R Hershey and Peder A Olsen. Approximating the kullback leibler divergence between gaussian mixture models. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages IV–317. IEEE, 2007.

[17] Mu Hu, Shuling Wang, Bin Li, Shiyu Ning, Li Fan, and Xiaojin Gong. Penet: Towards precise and efficient image guided depth completion. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13656–13662. IEEE, 2021.

[18] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 406–413, 2014.

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

[20] Kiriakos N Kutulakos and Steven M Seitz. A theory of shape by space carving. *International journal of computer vision*, 38(3):199–218, 2000.

[21] Maxime Lhuillier and Long Quan. A quasi-dense approach to surface reconstruction from uncalibrated images. *IEEE transactions on pattern analysis and machine intelligence*, 27(3):418–433, 2005.

[22] 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. *arXiv preprint arXiv:2205.14319*, 2022.

[23] 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.

[24] Keyang Luo, Tao Guan, Lili Ju, Haipeng Huang, and Yawei Luo. P-mvsnet: Learning patch-wise matching confidence aggregation for multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10452–10461, 2019.

[25] Keyang Luo, Tao Guan, Lili Ju, Yuesong Wang, Zhuo Chen, and Yawei Luo. Attention-aware multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1590–1599, 2020.

[26] Tao Luo, Zheng Ma, Zhiwei Wang, Zhiqin John Xu, and Yaoyu Zhang. An upper limit of decaying rate with respect to frequency in linear frequency principle model. In *Mathematical and Scientific Machine Learning*, pages 205–214. PMLR, 2022.

[27] 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.

[28] Zeyu Ma, Zachary Teed, and Jia Deng. Multiview stereo with cascaded epipolar raft. *arXiv preprint arXiv:2205.04502*, 2022.

[29] 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.

[30] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[31] Benjamin Peirce. Criterion for the rejection of doubtful observations. *The Astronomical Journal*, 2:161–163, 1852.

[32] 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.

[33] Charles M Rader and Bernard Gold. Digital filter design techniques in the frequency domain. *Proceedings of the IEEE*, 55(2):149–171, 1967.

[34] Douglas A Reynolds. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663), 2009.

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[36] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016.

[37] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European conference on computer vision*, pages 501–518. Springer, 2016.

[38] René Schuster, Oliver Wasenmuller, Christian Unger, and Didier Stricker. Ssgp: Sparse spatial guided propagation for robust and generic interpolation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 197–206, 2021.

[39] Steven M Seitz and Charles R Dyer. Photorealistic scene reconstruction by voxel coloring. *International Journal of Computer Vision*, 35(2):151–173, 1999.

[40] Duraisamy Sundararajan. *The discrete Fourier transform: theory, algorithms and applications*. World Scientific, 2001.

[41] SE Tavares. A comparison of integration and low-pass filtering. *IEEE Transactions on Instrumentation and Measurement*, 15(1/2):33–38, 1966.

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[43] 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.

[44] 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.

[45] Shaoqian Wang, Bo Li, and Yuchao Dai. Efficient multi-view stereo by iterative dynamic cost volume. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8655–8664, 2022.

[46] 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 *European Conference on Computer Vision*, pages 573–591. Springer, 2022.

[47] 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*, pages 6187–6196, 2021.

[48] Junhua Xi, Yifei Shi, Yijie Wang, Yulan Guo, and Kai Xu. Raymvsnet: Learning ray-based 1d implicit fields for accurate multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8595–8605, 2022.

[49] Hongbin Xu, Zhipeng Zhou, Yali Wang, Wenxiong Kang, Baigui Sun, Hao Li, and Yu Qiao. Digging into uncertainty in self-supervised multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6078–6087, 2021.

[50] 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.

[51] Qingshan Xu and Wenbing Tao. Learning inverse depth regression for multi-view stereo with correlation cost volume. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12508–12515, 2020.

[52] 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 *European conference on computer vision*, pages 674–689. Springer, 2020.

[53] 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.

[54] 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.

[55] Zhenpei Yang, Zhile Ren, Qi Shan, and Qixing Huang. Mvs2d: Efficient multi-view stereo via attention-driven 2d convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8574–8584, 2022.

[56] 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.

[57] 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.

[58] 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.

[59] Hongwei Yi, Zizhuang Wei, Mingyu Ding, Runze Zhang, Yisong Chen, Guoping Wang, and Yu-Wing Tai. Pyramid multi-view stereo net with self-adaptive view aggregation. In *European Conference on Computer Vision*, pages 766–782. Springer, 2020.

[60] 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.

[61] Zehao Yu and Shenghua Gao. Fast-mvsnet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1949–1958, 2020.

[62] Jingyang Zhang, Yao Yao, Shiwei Li, Zixin Luo, and Tian Fang. Visibility-aware multi-view stereo network. *arXiv preprint arXiv:2008.07928*, 2020.

[63] Xudong Zhang, Yutao Hu, Haochen Wang, Xianbin Cao, and Baochang Zhang. Long-range attention network for multi-view stereo. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3782–3791, 2021.

[64] Zixiang Zhao, Jiangshe Zhang, Shuang Xu, Zudi Lin, and Hanspeter Pfister. Discrete cosine transform network for guided depth map super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5697–5707, 2022.

[65] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*, 2018.