EC-MVSNet: Enhanced Cascaded Multi-View Stereo with Cross-Scale Relevance Integration

Shaoqian Wang[1,2], **Jiadai Sun**[3,4], **Bin Fan**[3], **Qiang Wang**[1,2], **Bin Lu**[1,2*], **Yuchao Dai**[3]

[1]Yanzhao Electric Power Laboratory of North China Electric Power University
[2]Hebei Key Laboratory of Knowledge Computing for Energy & Power
[3]School of Electronics and Information, Northwestern Polytechnical University
[4]Baidu Inc.
{wangshaoqian,52152569,lubin}@ncepu.edu.cn, {sunjiadai,binfan}@mail.nwpu.edu.cn, daiyuchao@gmail.com

Abstract

Cascade-based multi-scale architectures are currently the mainstream in Multi-view Stereo (MVS), achieving a balance between computational efficiency and reconstruction accuracy. However, existing cascade MVS methods suffer from significant limitations in cross-scale information utilization, where depth estimation processes operate independently across scales without fully exploiting the rich relevance between adjacent scales. To address this fundamental limitation, we propose an Enhanced Cascade Multi-View Stereo framework (EC-MVSNet), which introduces a novel cross-scale relevance integration strategy. Specifically, we introduce a Cross-Scale Feature-based Joint Construction (CFC) module to synergistically combine features from adjacent scales to build more reliable cost volumes. Additionally, a Cross-Scale Probability-guided Enhancement (CPE) module is proposed to propagate depth probability distributions across scales to guide cost volume enhancement. Furthermore, we propose a Monocular Feature-based Refinement (MFR) module to further enhance depth prediction accuracy by leveraging monocular priors. Extensive experiments demonstrate that EC-MVSNet achieves state-of-the-art performance on multiple benchmarks, validating the effectiveness of the cross-scale integration in improving MVS reconstruction quality.

Code — https://github.com/bdwsq1996/EC-MVSNet

Introduction

Multi-view stereo (MVS) stands as a classical and fundamental research direction in the field of computer vision, with broad applications spanning robotics, augmented reality, 3D modeling, and autonomous driving. In recent years, the rapid advancement of deep learning techniques have propelled learning-based MVS methods (Wang et al. 2024; Wang, Li, and Dai 2024; Cheng, Knoll, and Cao 2025; Kusupati et al. 2020) to the forefront of the field. These methods achieve leading performance (Cao, Ren, and Fu 2024; Wu et al. 2024) on multiple benchmarks compared to traditional MVS methods (Xu and Tao 2019; Galliani, Lasinger, and Schindler 2015; Schonberger and Frahm 2016; Locher, Perdoch, and Van Gool 2016) and demonstrate remarkable advantages in reconstructing scene details.

Learning-based MVS methods aim to predict the depth map for the reference image using a reference image and multiple source images within an end-to-end architecture. Following the pipeline proposed by MVSNet(Yao et al. 2018), most learning-based approaches first extract features from all images, followed by a depth estimation process consisting of cost volume construction, regularization, and depth prediction. Early learning-based methods (Yao et al. 2018; Chen et al. 2019; Xu and Tao 2020) typically performed depth estimation at a single scale, requiring substantial memory resources to construct a large-size cost volume, which limited their applicability to high-resolution input images. To address this limitation, CasMVSNet(Gu et al. 2020) introduced a cascade-based multi-scale framework that conducts depth estimation across multiple scales with progressively shrinking depth search ranges, achieving more efficient and accurate reconstruction results. These advantages have established cascade-based frameworks as the mainstream approach for learning-based MVS methods(Li et al. 2024; Zhang et al. 2023; Chen et al. 2020; Wang et al. 2025).

Although existing cascade-based MVS methods have achieved notable success, a critical limitation persists: the depth estimation process at each scale is executed independently with limited cross-scale interaction. While the framework leverages progressively refined depth ranges to guide subsequent stages, it fails to fully exploit the inherently rich cross-scale relevance information between adjacent scales. This limitation is evident in the neglected cross-scale relevance embedded in features and probability distributions, which constrains the potential of the cascade framework.

To address the above issues, we propose the Enhanced Cascade MVS framework (EC-MVSNet). As shown in Figure 1, compared to the general cascade-based framework, our method enables more integration of cross-scale relevance between adjacent scales. First, we design a Cross-Scale Feature-based Joint Construction (CFC) module to fully exploit the potential of cross-scale image features and construct the cost volume jointly, thereby improving the reliability of the cost volume. Next, a Cross-Scale Probability-guided Enhancement (CPE) module is intro-

*Corresponding author.

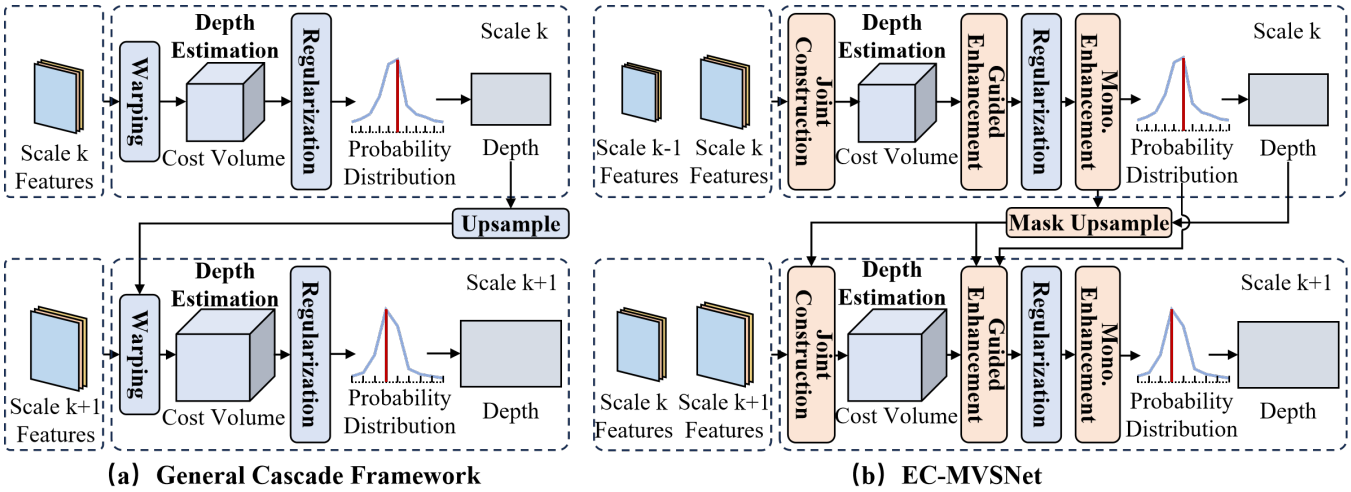

Figure 1: Visualization of the general cascade framework and proposed EC-MVSNet in cross-scale integration. The general cascade framework isolates adjacent scales, using upsampled predictions merely to define depth ranges. In contrast, EC-MVSNet enables joint cost volume construction from adjacent-scale features, enhances it via the probability distribution in the preceding scale, and propagates the optimized depth map through monocular prior-based mask upsampling.

duced, which analyzes the depth probability distribution at a lower-resolution scale to guide and enhance the cost volume at a subsequent finer scale. Furthermore, we propose a Monocular Feature-based Refinement (MFR) module that leverages monocular priors to improve depth estimation accuracy by refining the depth probability distribution and employing mask upsampling. Benefiting from the proposed cross-scale relevance integration strategy, EC-MVSNet achieves state-of-the-art (SOTA) performance on multiple benchmarks, demonstrating its effectiveness.

Our contributions are summarized as follows:

- We propose the CFC and CPE modules to explicitly integrate cross-scale relevance. The CFC module constructs cost volumes by jointly leveraging cross-scale features, while the CPE module enhances cost volume at the current scale by utilizing the depth probability distribution from the preceding scale as guidance.
- We propose an MFR module that leverages monocular priors to refine depth estimation through probability distribution optimization and mask-guided upsampling, propagating high-quality estimation to subsequent scales.
- Extensive experiments on DTU, Tanks and Temples, and ETH3D benchmarks demonstrate that our EC-MVSNet achieves state-of-the-art performance, validating the effectiveness of our cross-scale relevance integration strategy in addressing the limitations of existing cascade-based MVS frameworks.

Related Works

Learning-based Multi-view Stereo (MVS)

These methods have evolved from the foundational pipeline established by MVSNet (Yao et al. 2018), which consists of four key steps: feature extraction, cost volume construction, regularization, and depth map prediction. Early learning-based approaches primarily performed depth estimation at a single scale, necessitating the construction of a large cost volume to cover wide depth ranges, which severely limited their applicability to high-resolution reconstruction tasks due to excessive memory demands. To mitigate this limitation, the R-MVSNet series (Yao et al. 2019; Yan et al. 2020; Wei et al. 2021) introduced a sequential GRU-based regularization mechanism that processes the cost volume along the depth dimension in an unfolded manner, significantly reducing memory consumption but increasing computational time. In parallel, cascade-based methods (Gu et al. 2020; Wang et al. 2021; Ding et al. 2022; Wang et al. 2022; Liu et al. 2023; Cao, Ren, and Fu 2022; Fan et al. 2024; Mi, Di, and Xu 2022; Peng et al. 2022) pioneered a multi-scale framework that constructs cost volumes across progressively refined depth ranges at hierarchical scales, improving both computational efficiency and reconstruction accuracy.

Cross-Scale Relevance Integration in MVS

Existing cascade-based MVS methods primarily integrate cross-scale relevance by optimizing the contraction of depth search ranges across scales to enhance the depth prediction accuracy. UCS-Net (Cheng et al. 2020) adaptively determines the depth search range for the current scale by computing pixel-wise variance of the depth probability distribution from the preceding scale. EPP-MVSNet (Ma et al. 2021) proposes an entropy refining strategy that establishes the depth search range via entropy calculation of the previous scale's depth probability distribution. However, these methods lack integration of cross-scale correlations in other stages of the depth prediction pipeline, limiting the potential of cascade-based frameworks. In contrast, our method achieves comprehensive exploitation and integration of cross-scale correlations: the CFC module consoli-

dates representational advantages from multi-scale features, the CPE module leverages prior-scale probability distribution information to guide cost volume enhancement, and the MFR module further refines depth maps to provide reliable initial depth estimates for the subsequent scale.

Method

Overview

The proposed EC-MVSNet adopts a coarse-to-fine multi-scale framework to enhance depth estimation performance by explicitly integrating cross-scale relevance, as illustrated in Figure 2. In the initial scale, the depth estimation process follows the conventional multi-scale paradigm. In subsequent scales, two novel modules are introduced: the CFC module and CPE module, both operating on the cost volume. The CFC module jointly constructs the cost volume by leveraging features from adjacent scales, improving feature reliability. Meanwhile, the CPE module utilizes probability distributions inferred from the previous lower scales to guide and refine the cost volume at the current scale, enhancing its robustness. Following cost volume regularization, the proposed MFR module further optimizes the depth probability distribution, providing a more accurate depth map for subsequent scale through mask upsampling.

Cross-Scale Feature-based Joint Construction

Feature Extractor. The purpose of the CFC module is to further utilize the advantages of image features at adjacent scales, so that the constructed cost volume can better adapt to different types of object surface regions, thereby improving the performance of the algorithm in handling complex scene reconstruction tasks. To this end, multi-scale image features need to be extracted for the input reference image $\mathcal{I}_0$ and N-1 source images $\{\mathcal{I}_i\}_{i=1}^{N-1}$ respectively. Through a feature extractor based on the FPN network, the multi-scale image features $\{\mathcal{F}_i^k\}_{i=0}^{N-1}$ corresponding to each image are obtained, which include scales $k = 0, 1, 2, 3$. The size of the feature map at each scale is $\frac{H}{2^{3-k}} \times \frac{W}{2^{3-k}}$.

Joint Cost Volume Construction. The CFC module is designed to fully exploit the representational advantages of adjacent-scale features during cost volume construction, as illustrated in Figure 2. At the k-th scale, unlike previous cascade-based MVS methods that rely solely on features from the current scale to construct the cost volume, the CFC module innovatively employs features $\{\mathcal{F}_i^k\}_{i=0}^{N-1}$ and $\{\mathcal{F}_i^{k-1}\}_{i=0}^{N-1}$ to independently build cost volumes $\mathbf{C}'_k$ and $\mathbf{C}'_{k-1}$ under identical depth hypotheses $\{d_j^k\}_{j=1}^{\mathcal{Z}_k}$. These are subsequently fused to obtain the jointly optimized cost volume $\mathbf{C}_k$. Here, $\mathcal{Z}_k$ represents the number of depth hypotheses sampled in the depth search range $\boldsymbol{R}_k$ at the k-th scale. In this process, we design a sub-pixel feature matching strategy to address the dimensional mismatch issue when constructing the cost volume at scale k using features $\{\mathcal{F}_i^{k'}\}_{i=0}^{N-1}$ from arbitrary scale k'. Specifically, for a given pixel $\mathbf{p}_k$ in the reference feature map at scale k, it is first necessary to determine the corresponding pixel $\mathbf{p}_{k'}$ at scale k'. The detailed computational procedure is outlined as follows:

$$\mathbf{p}_{k'} = \mathbf{K}_{0,k'} \cdot \mathbf{K}_{0,k}^{-1} \cdot \mathbf{p}_k, \tag{1}$$

where $\mathbf{K}_{0,k'}$ and $\mathbf{K}_{0,k}$ respectively correspond to the intrinsic parameter matrices of the reference image camera at the scales k' and k. Subsequently, based on the depth hypothesis $\{d_j^k\}_{j=1}^{\mathcal{Z}_k}$, a series of corresponding pixel $\mathbf{p}_{i,j}^{k'}$ of $\mathbf{p}_{k'}$ in the source feature $\mathcal{F}_i^{k'}$ at the k'-th scale can be calculated with a differentiable homography transformation:

$$\mathbf{p}_{i,j}^{k'} = \mathbf{K}_{i,k'} \cdot (\mathbf{R}_{0,i} \cdot (\mathbf{K}_{0,k'}^{-1} \cdot \mathbf{p}_{k'} \cdot d_j^k) + \mathbf{t}_{0,i}), \tag{2}$$

where $\mathbf{K}_{i,k'}$ represents the intrinsic matrix of the i-th source feature at the k'-th scale. While $\mathbf{R}_{0,i}$ and $\mathbf{t}_{0,i}$ respectively represent the relative rotation matrix and translation vector between the reference image and i-th source image. It is worth noting that when k'=k, the homography transformation is consistent the existing cascade-based MVS works.

Subsequently, we compute the correlations and then fuse them to construction the cost volume $\mathbf{C}'_{k'}$. The correlation $\mathbf{c}_i^{k'}$ is calculated with reference feature $\mathcal{F}_0^{k'}(\mathbf{p})$ and the corresponding source feature $\mathcal{F}_i^{k'}(\mathbf{p}_{i,j}^{k'})$. Then, these correlations are weighted to construct the corresponding cost volume $\mathbf{C}'_{k'}$ as follows:

$$\mathbf{c}_i^{k'}(\mathbf{p}_k, d_j^k) = \langle \mathcal{F}_0^{k'}(\mathbf{p}_{k'}), \mathcal{F}_i^{k'}(\mathbf{p}_{i,j}^{k'}) \rangle_g, \tag{3}$$

$$\mathbf{C}'_{k'}(\mathbf{p}_k, d_j^k) = \frac{\sum_{i=1}^{N-1} \boldsymbol{w}_{i,k}(\mathbf{p}_k) \mathbf{c}_i^{k'}(\mathbf{p}_k, d_j^k)}{\sum_{i=1}^{N-1} \boldsymbol{w}_{i,k}(\mathbf{p}_k)}, \tag{4}$$

where $\boldsymbol{w}_{i,k}$ denotes pixel-wise weights (Wang et al. 2022), while $\langle \cdot, \cdot \rangle_g$ represents the group-wise correlation calculation. In CFC module, the cost volumes $\mathbf{C}'_{k-1}$ and $\mathbf{C}'_k$ are constructed with Eq. 4 based on the features $\{\mathcal{F}_i^{k-1}\}_{i=0}^{N-1}$ and $\{\mathcal{F}_i^k\}_{i=0}^{N-1}$. Subsequently, a series of 3D CNN layers is used to process these cost volumes respectively, and then aggregate them to obtain the cost volume $\mathbf{C}_k$. Notably, we have also explored alternative joint cost volume construction strategies, which are presented in the ablation study section.

Cross-Scale Probability-guided Enhancement

In the process of depth estimation, a regularization module is generally used to process the constructed cost volume, obtaining an initial probability distribution map, which is then refined by the proposed MFR module, denoted as $\mathcal{P}_k$. The probability distribution map reflects the probability that the depth of each pixel position p within the depth search range $\boldsymbol{R}_k$ at the current k-th scale corresponds to different depth hypotheses $\{d_j^k\}_{j=1}^{\mathcal{Z}_k}$. In the cascade framework, the depth search range will gradually shrink as the scale level increases, where the depth search range $\boldsymbol{R}_k$ consistently constitutes a subset of $\boldsymbol{R}_{k-1}$. Since $\mathcal{P}_{k-1}$ reflects the probability distribution within the entire $\boldsymbol{R}_{k-1}$, this means that by sampling the part corresponding to $\boldsymbol{R}_k$ in $\mathcal{P}_{k-1}$, the obtained local probability distribution can be a prior for the depth estimation within the $\boldsymbol{R}_k$. Therefore, we propose a CPE module to utilize the probability distribution map $\mathcal{P}_{k-1}$ as prior information to further guide and enhance the cost volume $\mathbf{C}_k$ at the scale k. Specifically, we analyze the relationship between the search ranges $\boldsymbol{R}_{k-1}$ and $\boldsymbol{R}_k$ to perform sampling, and subsequently leverage the sampled local probability distribution to perform guided enhancement on $\mathbf{C}_k$.

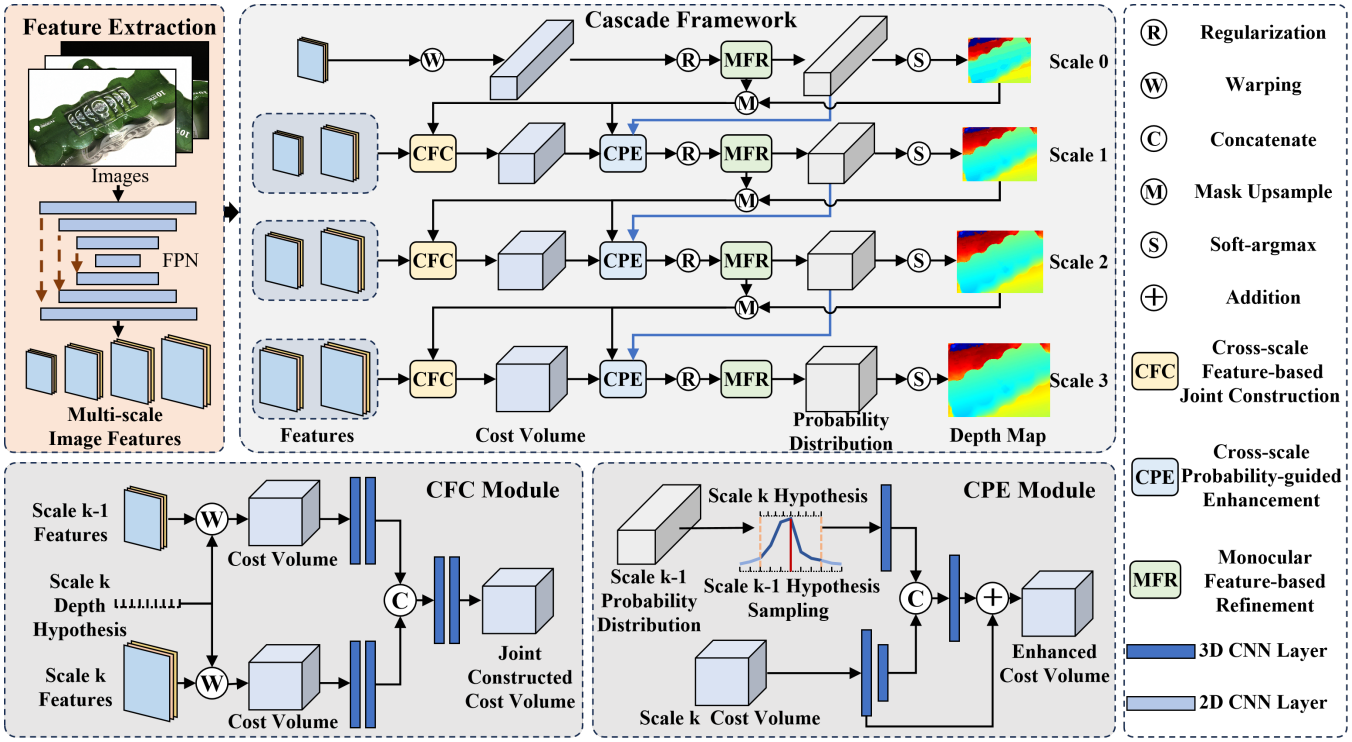

Figure 2: Overview of the proposed EC-MVSNet. Our method enhances reconstruction performance by extensively exploiting and integrating cross-scale relevance within a generic cascade framework. At finer scales (k>0), the CFC module is proposed to jointly construct cost volumes with adjacent-scale features. Subsequently, the CPE module enhances the cost volumes by utilizing the probability distribution from the preceding scale as guidance. After regularizing the cost volumes to establish depth probability distributions, a MFR module is utilized to incorporate monocular features for optimizing the depth probability distribution, further combining the mask upsampling strategy to propagate a reliable depth map to the subsequent scale.

Probability Distribution Sampling. As shown in Figure 3, we sampled $\mathcal{P}_{k-1}$ to obtain the corresponding local probability distribution $\mathcal{P}'_{k-1}$ for the range $\boldsymbol{R}_k$. Specifically, we first downsample the depth hypotheses $d_j^k \in \mathbb{R}^{\frac{H}{2^{3-k}} \times \frac{W}{2^{3-k}} \times Z_k}$ along the height and width dimensions to ensure consistency with the spatial dimensions of $\mathcal{P}_{k-1} \in \mathbb{R}^{\frac{H}{2^{4-k}} \times \frac{W}{2^{4-k}} \times Z_{k-1}}$. Subsequently, the local probability distribution $\mathcal{P}'_{k-1} \in \mathbb{R}^{\frac{H}{2^{4-k}} \times \frac{W}{2^{4-k}} \times Z_k}$ is constructed with linear interpolation on $\mathcal{P}_{k-1}$ based on the downsampled d_j^k.

Guided Enhancement. The CPE module employs a series of 3D CNN layers to separately process the cost volume $\mathbf{C}_k$ and the local probability distribution $\mathcal{P}'_{k-1}$, thereby achieving the enhancement effect. As shown in Figure 2, the module applies a 3D CNN layer with stride of 2 to downsample $\mathbf{C}_k$, ensuring that its spatial dimensions match those of $\mathcal{P}'_{k-1}$. This strategy enables the CPE module to effectively aggregate probability distribution information while maintaining computational efficiency. Subsequently, the processed $\mathcal{P}'_{k-1}$ and $\mathbf{C}_k$ are concatenated along the channel dimension and further refined through an additional 3D CNN layer. Finally, skip connections are employed to enhance the cost volume, which significantly improves the accuracy of depth estimation.

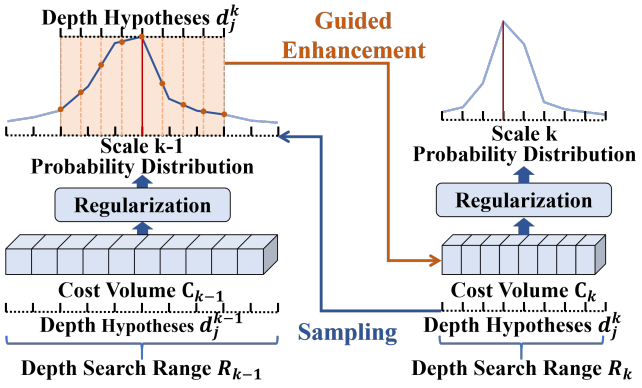

Figure 3: Visualization of the probability distribution sampling process in the CPE module. Based on the depth hypothesis d_j^k, linear interpolation is performed on the probability distribution $\mathcal{P}_{k-1}$ to extract a local probability distribution within the depth range $\boldsymbol{R}_k$, which is subsequently utilized to guide the enhancement of the cost volume $\mathbf{C}_k$.

Monocular Feature-based Refinement

To further enhance depth prediction accuracy, we propose the MFR module that introduces monocular features as pri-

ors to refine both the depth probability distribution and predicted depth maps. For this purpose, we leverage the DepthAnythingV2 (Yang et al. 2024) (ViT-b version) pre-trained model to process reference images and extend the extracted monocular features across multiple scales. These monocular features, rich in surface smoothness representations, effectively improve depth estimation consistency.

As illustrated in Figure 4, at the scale k, a series of 2D CNN layers process the concatenation of monocular features and the initial depth probability distribution from the regularization module to construct the refined depth probability distribution $\mathcal{P}_k$. Subsequently, the Soft-argmax function predicts depth values $\mathcal{D}_k$ from $\mathcal{P}_k$. Inspired by (Teed and Deng 2020), we employ a Mask Upsample strategy to simultaneously upsample and refine the depth map $\mathcal{D}_k \in \mathbb{R}^{\frac{H}{2^{3-k}} \times \frac{W}{2^{3-k}}}$, generating $\mathcal{D}_k^u \in \mathbb{R}^{\frac{H}{2^{2-k}} \times \frac{W}{2^{2-k}}}$ for propagation to the next scale. Specifically, we process monocular features and depth values $\mathcal{D}_k$ through 2D CNN layers and concatenation to construct a mask map, facilitating $\mathcal{D}_k$ upsampling and optimization to enhance depth continuity. Additionally, at scale 3, the mask sample strategy exclusively refines the depth map without upsampling, yielding the final depth prediction $\mathcal{D}_3^u$.

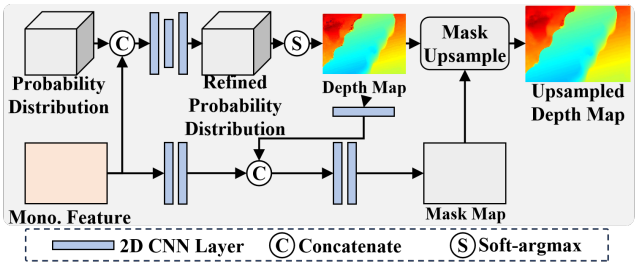

Figure 4: Visualization of the MFR module. The MFR module introduces monocular feature from the reference image to refine depth probability distributions and construct a mask with the estimated depth map, then propagates optimized depth estimates to the next scale through mask upsampling.

Loss Function

The proposed method employs cross-entropy to supervise the predicted probability distributions and L_1 loss to supervise mask upsampled depth, with the specific calculation defined as follows:

$$L = \sum_{k=0}^{3} \sum_{\mathbf{p} \in \Phi} -\mathcal{P}_k^{gt}(\mathbf{p}) log(\mathcal{P}_k(\mathbf{p})) + \|\mathcal{D}_{gt}(\mathbf{p}) - \mathcal{D}_k^u(\mathbf{p})\|_1, \tag{5}$$

where Φ denotes the set of all pixel positions with valid depth values in the ground truth depth map. $\mathcal{D}_{gt}$ denotes the ground truth depth with the same size as $\mathcal{D}_k^u$.

Experiments

Extensive experiments are conducted on DTU (Aanæs et al. 2016), Tanks and Temples (Knapitsch et al. 2017), and ETH3D (Schops et al. 2017) datasets to verify the effectiveness and superiority of EC-MVSNet. In addition, extensive ablation experiments are carried out on the DTU dataset.

Implementation Details

Training. The proposed method is implemented using the PyTorch framework, with training and fine-tuning conducted on two NVIDIA GeForce RTX 3090 GPUs. The Adam optimizer (Kingma and Ba 2014) is adopted, and the learning rate is dynamically adjusted using the OneCycleLR strategy (Smith and Topin 2019). The model is first trained 15 epochs on the DTU training dataset, and then evaluated on the DTU test dataset. Subsequently, an additional 10 epochs of fine-tuning are performed on the BlendedMVS dataset (Yao et al. 2020), followed by evaluations on the Tanks and Temples and ETH3D datasets. For training on the DTU dataset, the initial learning rate is set to 0.001, with $N = 5$ input images of resolution 640×512. During fine-tuning on the BlendedMVS dataset, the initial learning rate is 0.0004, with $N = 7$ input images at a resolution of 768×576. The number of depth hypotheses Z_k is set to 32, 16, 8, and 4 for scales $k = 0, 1, 2, 3$, respectively.

Evaluation. The proposed method is evaluated on the DTU, Tanks and Temples, and ETH3D datasets, with varying configurations. On the DTU dataset, $N = 5$ input images of resolution 1536×1152 are used. For the Tanks and Temples dataset, $N = 21$ input images at a resolution of 1920×1056 are employed. In the ETH3D experiments, $N = 16$ input images with a resolution of 2432×1600 are utilized. To remove outliers in the estimated depth maps, a reprojection strategy based on dynamic geometric consistency (Yan et al. 2020) is employed. All filtered depth maps are projected and then fused into a dense point cloud.

Metrics. The DTU employs distance-based metrics $[mm]$ to compute Accuracy, Completeness, and Overall, where lower values indicate better reconstruction quality. In contrast, Tanks and Temples and ETH3D use percentage-based metrics [%] to compute Precision, Recall, and F_1 scores, where higher values indicating superior performance.

Benchmark Performance

DTU. Quantitative comparisons between our method and other state-of-the-art cascade-based methods on the DTU dataset are presented in Table 1. Our method exhibits superior reconstruction performance on the DTU dataset. Compared with representative multi-scale methods (Wang et al. 2022; Ding et al. 2022; Chang et al. 2024; Liu et al. 2023), the proposed method shows better performance in three key metrics. Compared with the current state-of-the-art method MVSFormer++ (Cao, Ren, and Fu 2024), our method optimizes the *Overall* metric from 0.281 to 0.275, and optimizes the *Accuracy* metric is optimized from 0.309 to 0.297. The point clouds results are illustrated in Figure 5, our method demonstrates superior reconstruction performance at object edges and fine structural details compared to MVSFormer++, further validating the effectiveness of our method.

Tanks and Temples. The quantitative results of our method on the Intermediate and Advanced sets are compared with

Figure 5: Comparison of the point clouds reconstruction performance with MVSFormer++ (Cao, Ren, and Fu 2024) on multiple scenes in DTU benchmark. Our method achieves a more accurate reconstruction of detailed structures with fewer outliers.

Methods	Overall↓	Acc. ↓	Comp. ↓
PatchmatchNet CVPR 21	0.352	0.427	0.277
CasMVSNet CVPR 20	0.348	0.346	0.351
RayMVSNet CVPR 22	0.330	0.341	0.319
Effi-MVS CVPR 22	0.317	0.321	0.313
MVSTER ECCV 22	0.303	0.340	0.266
EI-MVSNet TIP 24	0.303	0.346	0.260
GeoMVSNet CVPR 23	0.295	0.331	0.259
ET-MVSNet CVPR 23	0.291	0.329	0.253
MVSFormer TMLR 22	0.289	0.327	0.251
GoMVS CVPR 24	0.287	0.347	**0.227**
RRT-MVS AAAI 25	0.285	0.309	0.261
MVSFormer++ ICLR 24	0.281	0.309	0.252
Ours	**0.275**	**0.297**	0.253

Table 1: Performance on DTU dataset. Best result in bold.

those of existing methods in Table 2. The proposed EC-MVSNet achieves the state-of-the-art performance in terms of the mean F_1 score. Additionally, our method achieves optimal or near-optimal performance in most scenarios. Compared with the previous best cascade-based method RRT-MVS (Jiang et al. 2025), the mean F_1 score of the proposed method in the Intermediate set is improved from 68.16 to 69.32, and that in the Advanced set is improved from 43.29 to 44.63. These results further prove the effectiveness and superiority of the proposed method. Furthermore, Figure 6 shows a qualitative comparison of the reconstruction results of the proposed method and other SOTA cascade-based works for the M60 and Temple scenes. These result images are sourced from online public evaluations. The darker the color of the scene area, the larger the point cloud distance error. It can be observed that the point cloud results obtained by the proposed method have fewer outliers, demonstrating excellent accuracy of the reconstruction results.

ETH3D. The quantitative results on the two Training and Testing sets of ETH3D dataset are shown in Table 3. Compared with existing cascade-based methods, the proposed method achieves state-of-the-art performance in terms of F_1 score and Precision score, and comprehensive performance

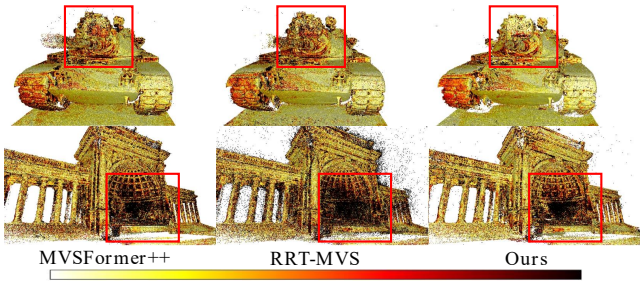

Figure 6: Comparison of Precision performance with state-of-the-art cascade-based works (Cao, Ren, and Fu 2024; Jiang et al. 2025) on the Tanks and Temples benchmark. Darker regions indicate larger errors.

for the Recall score on both sets.

Ablation Study

Effectiveness Analysis of the CFC Module and CPE Module. This subsection adopts the standard coarse-to-fine multi-scale framework as the basic method and gradually introduces the proposed CFC, CPE and MFR modules to specifically analyze their impact on the framework performance. The results of the comparative experiments are shown in Table 4. When the three proposed modules are used individually, the CPE module demonstrates the lowest overall error of 0.284, proving that integrating cross-scale depth probability distribution relevance plays a crucial role in improving reconstruction quality. When the CFC and CPE modules are jointly employed, our method achieves comprehensive integration of cross-scale relevance, reducing the overall error from 0.303 to 0.279. Building upon this, the introduction of the MFR module for depth prediction optimization effectively enhances reconstruction quality, improving the overall score to 0.275.

Multi-Scale Feature Information Aggregation Analysis. In the CFC module, cost volumes are constructed using image features from both the current scale and the previous scale to achieve aggregation of multi-scale feature information. We provide a detailed analysis of different feature aggregation methods across various hierarchy levels and num-

Methods	Intermediate ↑									Advanced ↑						
	Mean	Fam.	Franc.	Horse	L.H.	M60	Pan.	P.G.	Train	**Mean**	Audi.	Ballr.	Courtr.	Mus.	Palace	Temple
PatchmatchNet CVPR 21	53.15	66.99	52.64	43.24	54.87	52.87	49.54	54.21	50.81	32.31	23.69	37.73	30.04	41.80	28.31	32.29
CasMVSNet CVPR 20	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
MVSTER ECCV 22	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
EPP-MVSNet ICCV 21	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
ET-MVSNet CVPR 23	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
EIA-MVSNet RAL 24	65.67	81.67	70.17	55.17	66.21	64.29	64.64	63.95	59.27	41.83	29.95	47.36	40.19	52.49	36.17	44.82
RA-MVSNet CVPR 23	65.72	82.44	66.61	58.40	64.78	**67.14**	65.60	62.74	58.08	39.93	29.14	46.04	40.30	53.22	34.63	36.28
DS-PMNet AAAI 24	64.16	81.11	63.43	60.84	62.23	64.96	61.92	61.41	57.35	39.78	28.52	44.93	39.12	51.68	33.77	40.67
GoMVS CVPR 24	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
MVSFormer TMLR 22	66.41	82.06	69.33	60.59	68.54	65.67	64.07	61.45	59.54	40.87	28.22	46.75	39.30	52.88	35.16	42.95
MVSFormer++ ICLR 24	67.18	**82.69**	69.44	64.24	69.16	64.13	66.43	61.19	60.12	41.60	29.93	45.69	39.46	53.58	35.56	45.39
RRT-MVS AAAI 25	68.16	82.54	72.31	61.44	69.89	65.32	68.88	64.45	60.48	43.29	30.95	46.42	41.13	55.46	**37.63**	48.12
Ours	**69.32**	82.19	**73.20**	65.55	**69.98**	66.44	**68.99**	**66.10**	**62.07**	**44.63**	**33.30**	**49.28**	41.02	**55.90**	37.39	**50.89**

Table 2: Performance on Tanks and Temples dataset. Best result in bold.

Methods	Training ↑			Testing ↑		
	F_1	Pre.	Recall	F_1	Pre.	Recall
PatchmatchNet CVPR 21	64.21	65.43	64.81	73.12	69.71	77.46
GBi-Net CVPR 22	70.78	73.17	69.21	78.40	82.02	75.65
MVSTER ECCV 22	72.06	68.08	76.92	79.01	77.09	82.47
EPP-MVSNet ICCV 21	74.00	82.76	67.58	83.40	85.47	81.79
EIA-MVSNet RAL 24	75.64	78.57	73.48	83.31	82.19	84.84
GoMVS CVPR 2024	79.16	81.22	**77.65**	85.91	85.50	**86.85**
Ours	**79.44**	**83.70**	76.05	**86.01**	**88.01**	84.53

Table 3: Performance results on ETH3D. Best result in bold.

CFC	CPE	MFR	Overall↓	Acc.↓	Comp.↓
–	–	–	0.303	0.323	0.273
✓	–	–	0.290	0.314	0.266
–	✓	–	0.284	0.309	0.259
✓	✓	–	0.279	0.304	0.254
–	–	✓	0.288	0.308	0.268
✓	✓	✓	**0.275**	**0.297**	**0.253**

Table 4: Ablation study for the proposed components.

bers of scales, evaluating three aspects: reconstruction performance, time (s), and GPU memory usage (GB). Specifically, we examined four aggregation approaches: (1) aggregating current scale features with adjacent lower scale features; (2) aggregating with adjacent higher scale features; (3) aggregating with both adjacent higher and lower scale features; and (4) aggregating features from all scales. The experimental results are presented in Table 5. The results show that the method using adjacent higher-scale features achieves the lowest runtime and GPU memory consumption, as it eliminates the need for CFC module-based information aggregation at the highest resolution scale, thereby reducing computational overhead. Compared to our adopted method of using adjacent lower-scale features, the methods that aggregate with both higher and lower scales or all scales show marginal improvements in reconstruction quality but significantly increase both runtime and memory usage. Notably, the full-scale feature aggregation method incurs the highest additional costs, requiring the simultaneous construction of four cost volumes at high resolution scales, which results in an extra 0.36 seconds of processing time and 4.0GB of additional GPU memory consumption.

Method	Overall	Acc.	Comp.	Time	Mem.
with low	0.275	0.297	0.253	0.44	5.4
with high	0.278	0.301	0.255	**0.37**	**4.7**
low and high	0.274	0.298	0.250	0.48	6.5
all	**0.274**	**0.297**	**0.250**	0.80	9.4

Table 5: Ablation study of the CFC module.

Conclusion

In this paper, we present EC-MVSNet, an enhanced cascade MVS framework that addresses critical limitations in existing methods by enabling comprehensive cross-scale relevance integration through three synergistic modules. The CFC module jointly constructs cost volumes by leveraging cross-scale image features, enhancing their reliability. The CPE module enhances cost volumes using depth probability distributions from coarser scales. The MFR module further improves accuracy via depth probability distribution optimization and mask upsampling, propagating high-quality depth estimates across scales. Benefiting from this cross-scale relevance integration strategy, EC-MVSNet achieves SOTA performance on multiple benchmarks.

Acknowledgements

This work is partly supported by the National Natural Science Foundation of China (62501236), the Beijing Natural Science Foundation (4254105), and the Hebei Province Natural Science Foundation (F2025502023).

References

Aanæs, H.; Jensen, R. R.; Vogiatzis, G.; Tola, E.; and Dahl, A. B. 2016. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120(2): 153–168.

Cao, C.; Ren, X.; and Fu, Y. 2022. MVSFormer: Multi-View Stereo by Learning Robust Image Features and Temperature-based Depth. *Transactions on Machine Learning Research*, 158–1608.

Cao, C.; Ren, X.; and Fu, Y. 2024. MVSFormer++: Revealing the Devil in Transformer's Details for Multi-View Stereo. In *Proceedings of the International Conference on Learning Representations*, 1–14.

Chang, J.; He, J.; Zhang, T.; Yu, J.; and Wu, F. 2024. EI-MVSNet: Epipolar-Guided Multi-View Stereo Network with Interval-Aware Label. *IEEE Transactions on Image Processing*, 753 – 766.

Chen, R.; Han, S.; Xu, J.; and Su, H. 2019. Point-based multi-view stereo network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1538–1547.

Chen, R.; Han, S.; Xu, J.; and Su, H. 2020. Visibility-aware point-based multi-view stereo network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10): 3695–3708.

Cheng, S.; Xu, Z.; Zhu, S.; Li, Z.; Li, L. E.; Ramamoorthi, R.; and Su, H. 2020. Deep stereo using adaptive thin volume representation with uncertainty awareness. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2524–2534.

Cheng, Y.; Knoll, A.; and Cao, H. 2025. UR-Net: uncertainty-aware refinement network for event-based stereo depth estimation. *Visual Intelligence*, 3(1): 18.

Ding, Y.; Yuan, W.; Zhu, Q.; Zhang, H.; Liu, X.; Wang, Y.; and Liu, X. 2022. Transmvsnet: Global context-aware multi-view stereo network with transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8585–8594.

Fan, B.; Dai, Y.; Seo, Y.; and He, M. 2024. A revisit of the normalized eight-point algorithm and a self-supervised deep solution. *Visual Intelligence*, 2(1): 3.

Galliani, S.; Lasinger, K.; and Schindler, K. 2015. Massively parallel multiview stereopsis by surface normal diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 873–881.

Gu, X.; Fan, Z.; Zhu, S.; Dai, Z.; Tan, F.; and Tan, P. 2020. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2495–2504.

Jiang, J.; Wang, L.; Yu, H.; Hu, T.; Chen, J.; and Ma, H. 2025. RRT-MVS: Recurrent Regularization Transformer for Multi-View Stereo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 3994–4002.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization.

Knapitsch, A.; Park, J.; Zhou, Q.-Y.; and Koltun, V. 2017. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4): 1–13.

Kusupati, U.; Cheng, S.; Chen, R.; and Su, H. 2020. Normal assisted stereo depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2189–2199.

Li, H.; Guo, Y.; Zheng, X.; and Xiong, H. 2024. Learning deformable hypothesis sampling for accurate patchmatch multi-view stereo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 3082–3090.

Liu, T.; Ye, X.; Zhao, W.; Pan, Z.; Shi, M.; and Cao, Z. 2023. When epipolar constraint meets non-local operators in multi-view stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 18088–18097.

Locher, A.; Perdoch, M.; and Van Gool, L. 2016. Progressive prioritized multi-view stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3244–3252.

Ma, X.; Gong, Y.; Wang, Q.; Huang, J.; Chen, L.; and Yu, F. 2021. EPP-MVSNet: Epipolar-Assembling Based Depth Prediction for Multi-View Stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5732–5740.

Mi, Z.; Di, C.; and Xu, D. 2022. Generalized binary search network for highly-efficient multi-view stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 12991–13000.

Peng, R.; Wang, R.; Wang, Z.; Lai, Y.; and Wang, R. 2022. Rethinking depth estimation for multi-view stereo: A unified representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8645–8654.

Schonberger, J. L.; and Frahm, J.-M. 2016. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4104–4113.

Schops, T.; Schonberger, J. L.; Galliani, S.; Sattler, T.; Schindler, K.; Pollefeys, M.; and Geiger, A. 2017. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3260–3269.

Smith, L. N.; and Topin, N. 2019. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, 369–386.

Teed, Z.; and Deng, J. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *Proceedings of the European Conference on Computer Vision*, 402–419.

Wang, F.; Galliani, S.; Vogel, C.; Speciale, P.; and Pollefeys, M. 2021. PatchmatchNet: Learned Multi-View Patchmatch Stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 14194–14203.

Wang, S.; Ding, X.; Mao, Y.; and Dai, Y. 2025. ETV-MVS: Robust Visibility-Aware Multi-View Stereo with Epipolar Line-Based Transformer. *Big Data Mining and Analytics*, 8(3): 520–533.

Wang, S.; Li, B.; and Dai, Y. 2024. Efficient multi-view stereo by dynamic cost volume and cross-scale propagation. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(10): 9414–9427.

Wang, S.; Li, B.; Yang, J.; and Dai, Y. 2024. Adaptive Feature Enhanced Multi-View Stereo With Epipolar Line Information Aggregation. *IEEE Robotics and Automation Letters*, 9(11): 10439–10446.

Wang, X.; Zhu, Z.; Huang, G.; Qin, F.; Ye, Y.; He, Y.; Chi, X.; and Wang, X. 2022. MVSTER: Epipolar transformer for efficient multi-view stereo. In *Proceedings of the European Conference on Computer Vision*, 573–591.

Wei, Z.; Zhu, Q.; Min, C.; Chen, Y.; and Wang, G. 2021. AA-RMVSNet: Adaptive Aggregation Recurrent Multi-view Stereo Network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6187–6196.

Wu, J.; Li, R.; Xu, H.; Zhao, W.; Zhu, Y.; Sun, J.; and Zhang, Y. 2024. Gomvs: Geometrically consistent cost aggregation for multi-view stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 20207–20216.

Xu, Q.; and Tao, W. 2019. Multi-scale geometric consistency guided multi-view stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5483–5492.

Xu, Q.; and Tao, W. 2020. Learning inverse depth regression for multi-view stereo with correlation cost volume. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 12508–12515.

Yan, J.; Wei, Z.; Yi, H.; Ding, M.; Zhang, R.; Chen, Y.; Wang, G.; and Tai, Y.-W. 2020. Dense hybrid recurrent multi-view stereo net with dynamic consistency checking. In *Proceedings of the European Conference on Computer Vision*, 674–689.

Yang, L.; Kang, B.; Huang, Z.; Zhao, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024. Depth anything v2. *Proceedings of the Advances in Neural Information Processing Systems*, 37: 21875–21911.

Yao, Y.; Luo, Z.; Li, S.; Fang, T.; and Quan, L. 2018. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision*, 767–783.

Yao, Y.; Luo, Z.; Li, S.; Shen, T.; Fang, T.; and Quan, L. 2019. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5525–5534.

Yao, Y.; Luo, Z.; Li, S.; Zhang, J.; Ren, Y.; Zhou, L.; Fang, T.; and Quan, L. 2020. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1790–1799.

Zhang, Z.; Peng, R.; Hu, Y.; and Wang, R. 2023. Geomvsnet: Learning multi-view stereo with geometry perception. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 21508–21518.