

MonoMVSNet: Monocular Priors Guided Multi-View Stereo Network

Jianfei Jiang, Qiankun Liu*, Haochen Yu, Hongyuan Liu, Liyong Wang, Jiansheng Chen, Huimin Ma*
University of Science and Technology Beijing, China

{jiangjf, haochen.yu, hongyuanliu, wangly}@xs.ustb.edu.cn, {liuqk3, jschen, mhmpub}@ustb.edu.cn

Abstract

Learning-based Multi-View Stereo (MVS) methods aim to predict depth maps for a sequence of calibrated images to recover dense point clouds. However, existing MVS methods often struggle with challenging regions, such as texture-less regions and reflective surfaces, where feature matching fails. In contrast, monocular depth estimation inherently does not require feature matching, allowing it to achieve robust relative depth estimation in these regions. To bridge this gap, we propose MonoMVSNet, a novel monocular feature and depth guided MVS network that integrates powerful priors from a monocular foundation model into multi-view geometry. Firstly, the monocular feature of the reference view is integrated into source view features by the attention mechanism with a newly designed cross-view position encoding. Then, the monocular depth of the reference view is aligned to dynamically update the depth candidates for edge regions during the sampling procedure. Finally, a relative consistency loss is further designed based on the monocular depth to supervise the depth prediction. Extensive experiments demonstrate that MonoMVSNet achieves state-of-the-art performance on the DTU and Tanks-and-Temples datasets, ranking first on the Tanks-and-Temples Intermediate and Advanced benchmarks. The source code is available at <https://github.com/JianfeiJ/MonoMVSNet>.

1. Introduction

Multi-View Stereo (MVS) is a fundamental task in computer vision that aims to reconstruct dense 3D geometry [19, 45, 51] from a set of calibrated images. While traditional MVS methods [8, 34, 36, 46–48] have made significant progress, they remain limited by handcrafted feature representations. The rapid advancement of deep learning in recent years has driven the development of learning-based MVS [9, 12, 13, 41, 42], which leverage the powerful representation capabilities of deep neural networks, leading to substantial performance gains over traditional approaches.

*Corresponding author

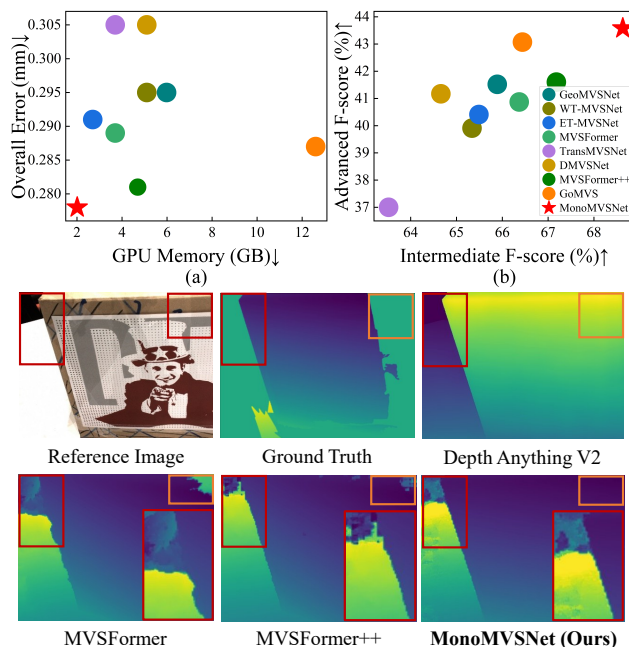


Figure 1. **Row 1:** (a) Comparison with SOTA methods in terms of overall error and GPU memory consumption at a resolution of 832×1152 with 5-view images on the DTU [1] test set, where **lower is better**; (b) Comparison with SOTA methods on the Tanks-and-Temples [16] benchmark, where **higher is better**. **Row 2-3:** Qualitative depth comparison with Depth Anything V2 [40], MVSFormer [2], and MVSFormer++ [3] on scan13 from DTU test set. Our method produces more accurate metric depth in edge (red bounding box) and textureless (orange bounding box) regions.

MVS can essentially be understood as a one-to-many feature matching problem [7]. However, feature matching is often challenged by difficult regions, such as texture-less areas, reflective surfaces, and depth discontinuity edges, leading to suboptimal reconstruction results, as shown in Fig. 1. To mitigate this issue, recent MVS methods have focused on enhancing feature extraction, as improved features enable more effective feature matching. The conventional approach for feature extraction involves using a Feature Pyramid Network (FPN) [14] to extract multi-scale features, facilitating the construction of multi-level cost volumes for

coarse-to-fine depth prediction. Building on this, recent methods have integrated deformable convolutions [6, 7, 21], attention mechanisms [7, 18, 20], and pre-trained Vision Transformers (ViTs) [2, 3, 29] to extract more reliable feature representations.

On the other hand, monocular depth estimation has seen significant progress in recent years [39, 40]. Unlike multi-view depth estimation, monocular depth models leverage neural networks to learn contextual cues for depth prediction, inherently mitigating mismatches in challenging regions. Recent monocular foundation models trained on large-scale real-world datasets exhibit strong zero-shot generalization capabilities. However, pre-trained monocular foundation models typically provide only relative depth and lack the geometric constraints of multi-view stereo, limiting their utility in downstream tasks. As illustrated in Fig. 1 (row 2), while the monocular depth estimation model (i.e., Depth Anything V2 [40]) generates visually compelling results, a significant scale ambiguity persists between the predicted depth and the ground truth.

To address this, we propose MonoMVSNet, which integrates the powerful priors from the monocular foundation model with multi-view geometry to construct a more robust and stronger MVS network. Although previous works [2, 3] have explored the usage of pre-trained ViTs to improve feature representations in multi-view stereo, they rely on complex training strategies and architectures that require pre-trained ViT features for all input views, introducing significant overhead. In contrast, our approach uses a simpler architecture, achieving better performance with lower GPU memory consumption, as shown in Fig. 1.

To fully exploit the strong generalization ability of the monocular foundation model, we first utilize pre-trained monocular features for robust feature extraction. Specifically, we extract monocular features only from the reference view and combine them with FPN features. Subsequently, for the reason that traditional positional encoding is designed for 2D images and does not capture 3D spatial relationships across viewpoints, we propose a novel Cross-View Position Encoding (CVPE) tailored for MVS to enhance both intra-view and inter-view attention mechanisms, effectively improving the representational capacity of source features. Furthermore, monocular depth provides fine-grained relative depth information in edge regions, which is essential for capturing depth discontinuities. To maximize the utility of this information, we introduce a monocular depth alignment module that aligns monocular relative depth with the predicted depth and guides the dynamic depth sampling process, ensuring a more accurate selection of depth candidates. Finally, we propose a relative consistency loss to enforce consistency between the aligned monocular depth and the predicted depth.

In summary, our contributions are as follows:

- We design a simple and effective approach that utilizes monocular feature priors to construct a powerful feature extractor. The proposed cross-view position encoding significantly improves the efficiency of feature exchange between different views.
- We design a dynamic depth sampling strategy using monocular relative depth priors and enforce relative depth consistency through a relative consistency loss, thereby improving the representation of depth discontinuities.
- MonoMVSNet achieves state-of-the-art performance on the DTU dataset and the Tanks-and-Temples benchmark.

2. Related Work

Learning-based Multi-View Stereo (MVS). MVS aims to reconstruct dense 3D representations from multiple images captured from different viewpoints. MVSNet [41], the first end-to-end learning-based MVS method, consists of four main steps: feature extraction, cost volume construction, cost volume regularization, and depth prediction. However, the high memory demand of 3D CNNs for cost volume regularization remains a limitation. Subsequent methods [32, 37, 42] reduced memory demands with RNN-based regularization, while cascade-based approaches [4, 9, 38] improved efficiency via coarse-to-fine processing. However, existing methods still struggle with reconstruction quality, emphasizing the need for better feature representations and depth sampling strategies.

Feature Representation in MVS. Effective feature representation is crucial for MVS performance. Cascade-based MVS methods [9] leverage FPN [14] for multi-scale feature extraction. To further improve this, some studies [7, 18, 20] incorporate Transformers [29] to aggregate global features through intra- and inter-view attentions. MVSFormer [2] fine-tunes a pre-trained Vision Transformer (ViT) [5] on high-resolution images for better feature representations. Building on this, MVSFormer++ [3] further improves feature learning by injecting cross-view information into the pre-trained DINOv2 [22]. In contrast to these methods, we integrate monocular features from pre-trained monocular foundation models [39, 40] with FPN features for the reference view only, and transfer these features to source views using cross-view position encoding to efficiently enhance feature representation robustness, while greatly reducing overhead.

Depth Sampling in MVS. In addition to feature extraction, depth sampling strategies are crucial for constructing accurate cost volumes in MVS methods. These methods construct the cost volume by sampling multiple depth candidates across the depth range and warping source feature maps onto corresponding depth hypothesis planes in the reference view. MVSNet [41] performs dense depth sampling, which is inefficient. CasMVSNet [9] introduces a coarse-to-fine depth sampling strategy. Building on this strategy, later

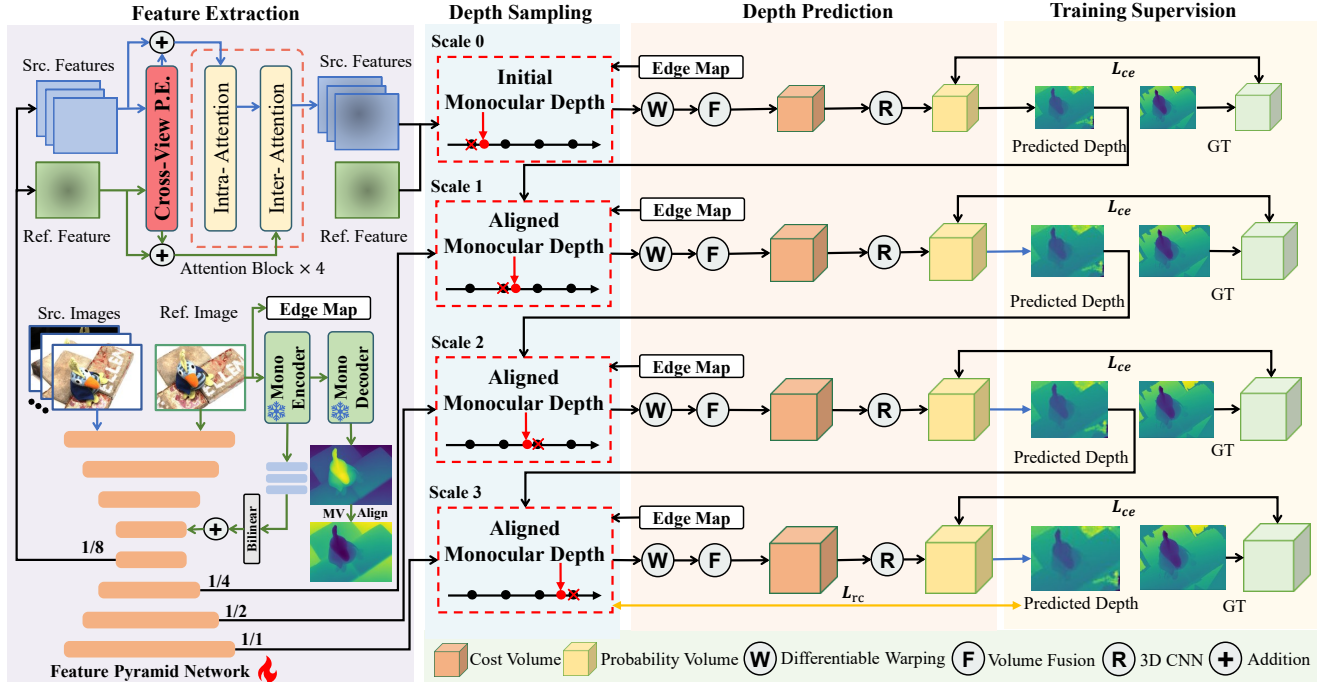


Figure 2. **Overview of the proposed MonoMVSNet.** (1) Exploitation of monocular feature: The reference monocular feature, extracted by the mono encoder model, is used to enhance the reference FPN feature and integrated into source features by attention mechanism with cross-view position encoding. (2) Exploitation of monocular depth: The monocular depth of the reference image, output by the mono decoder, is aligned to guide the depth sampling procedure and supervise the predicted depth with relative consistency loss.

studies have to further enhance memory efficiency and inference speed. For instance, IS-MVSNet [30] incorporates an importance sampling module, GBi-Net [21] employs a generalized binary search strategy, and MaGNet [21] samples based on the monocular depth probability distribution. These methods focus on sampling fewer depth candidates to improve memory efficiency and inference speed. In contrast, we utilize monocular depth to dynamically guide the depth sampling process, obtaining more reasonable depth candidates for better depth estimation, and thereby enhancing reconstruction performance.

Monocular Depth Estimation (MDE). Despite improvements in MVS methods, accurately reconstructing regions such as textureless areas, depth discontinuities, and reflective surfaces remains challenging. Recently, monocular depth estimation has made substantial progress. MiDaS [25] pioneered zero-shot generalization, while DPT [26] leveraged Transformers for fine-grained estimation. Depth Anything V1 [39] leverages a large number of unlabeled images, overcoming the traditional issue of insufficient labeled data. Depth Anything V2 [40] further incorporates knowledge distillation, allowing for robust depth prediction across complex scenarios. However, the depth maps generated by these monocular depth estimation methods often suffer from scale ambiguity, making them unsuitable for di-

rect use in downstream tasks (e.g., 3D reconstruction). To address this, we combine the strengths of monocular depth estimation and multi-view stereo to achieve robust multi-view depth estimation in challenging regions.

3. Methodology

3.1. Overview

The framework of MonoMVSNet is depicted in Fig. 2, which efficiently integrates monocular features and depth from a pre-trained monocular foundation model [40]. For the usage of monocular feature, we only feed the monocular model with the reference image to get the reference monocular feature, which avoids the overhead introduced by the monocular model as much as possible and also boosts the model performance. The reference monocular feature is integrated into source features through the attention mechanism, which is enhanced by the newly designed cross-view position encoding (Section 3.2). For the usage of monocular depth, we first align it with the coarse depth produced by MonoMVSNet by filtering unreliable positions. The aligned monocular depth is used to replace the depth candidates around the edge regions in the depth sampling procedure (Section 3.3) and to supervise the depth produced by MonoMVSNet with relative consistency loss (Section 3.5).

3.2. Monocular Feature for Feature Extraction

Given N input images $\{\mathbf{I}_n\}_{n=0}^{N-1} \in \mathbb{R}^{3 \times H \times W}$, consisting of a reference image (denoted as \mathbf{I}_0) and $N - 1$ source images, our goal is to estimate a depth map for the reference image using these input images and their corresponding camera parameters. Here, H and W represent the height and width of the input images, respectively.

Feature Extraction for Different Images. We employ a 4-layer FPN to extract multi-scale FPN features for each of the source images. Let s be the scale index and C be the channel dimensionality of FPN, the FPN feature of the $N - 1$ source images can be denoted as $\{\mathbf{F}_{n,s} \in \mathbb{R}^{C \times \frac{H}{2^{3-s}} \times \frac{W}{2^{3-s}}} | s = 0, 1, 2, 3\}_{n=1}^{N-1}$.

To extract features for the reference image, we first feed it to the pre-trained monocular model (specifically, Depth Anything V2 [40]), which produces the reference monocular feature $\mathbf{F}_0^{mono} \in \mathbb{R}^{C' \times h \times w}$, where C' is the channel dimensionality of the pre-trained monocular model and (h, w) is the spatial resolution of the feature. At the same time, the reference image is also fed into the encoder of FPN, producing the feature map of the lowest resolution, which is denoted as $\mathbf{F}_0^{enc} \in \mathbb{R}^{C \times \frac{H}{8} \times \frac{W}{8}}$. The feature \mathbf{F}_0^{mono} is processed by a convolution layer and a bilinear upsampling operation to align the channel dimensionality and spatial resolution with \mathbf{F}_0^{enc} , which are added with each other:

$$\mathbf{F}_0 = \mathbf{F}_0^{enc} \oplus \text{Bilinear}(\text{Conv}(\mathbf{F}_0^{mono})). \quad (1)$$

The feature \mathbf{F}_0 is further fed into the FPN decoder to get the multi-scale features for the reference image \mathbf{I}_0 , which is denoted as $\{\hat{\mathbf{F}}_{0,s} \in \mathbb{R}^{C \times \frac{H}{2^{3-s}} \times \frac{W}{2^{3-s}}} | s = 0, 1, 2, 3\}$.

In our design, the overhead introduced by the pre-trained monocular model is reduced as much as possible since only the reference image is processed by the monocular model. However, only the reference feature contains priors from the monocular model. To solve this, we use the attention mechanism to enhance source features with a newly designed cross-view position encoding, which also integrates the monocular prior from the reference feature to source features. As we will see in Section 4.4, such design can also boost the overall performance of MonoMVSNet compared to the peer model design that extracts and integrates the monocular feature for all source images.

Cross-View Position Encoding for Attention. Exploiting the intra-view and inter-view attention to enhance source features is a common practice in existing MVS works [7, 18, 20]. Though the source features can be enhanced to some extent in these works, they usually use the relative or absolute position encoding for attention, neglecting the importance of 3D spatial information across different views. To alleviate this, we propose a novel Cross-View Position Encoding (CVPE) to strengthen intra-view and inter-view attention interactions, which integrates the priors from the

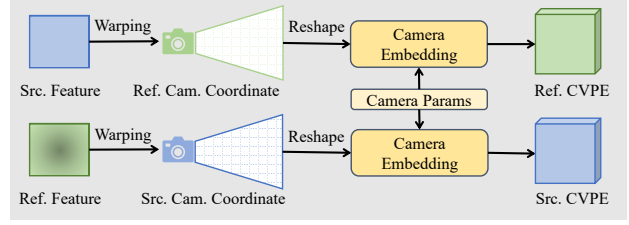


Figure 3. **Illustration of the Cross-View Position Encoding (CVPE).** Each pair of reference and source features is warped into the respective views. Together with the camera parameters, they undergo camera embedding to generate the CVPE for both the reference and source features.

reference feature to source features more effectively.

Supposing the depth hypotheses in the 0-th scale are $\{d_{i,0}\}_{i=0}^{D_0-1}$, the n -th source feature $\mathbf{F}_{n,0}$ is warped to these D_0 hypothesis planes using camera intrinsic matrix \mathbf{K}_n of the n -th source view, rotation matrix $\mathbf{R}_{n \rightarrow 0}$ and translation vector $\mathbf{t}_{n \rightarrow 0}$ from the n -th source view to the reference view. We denote the warped feature as $\mathbf{F}_{n \rightarrow 0,0} \in \mathbb{R}^{C \times D_0 \times \frac{H}{8} \times \frac{W}{8}}$. Given the 2D coordinate $\mathbf{c}_{n,0}$ in the n -th source feature $\mathbf{F}_{n,0}$, the warped 2D coordinate $\mathbf{c}_{n \rightarrow 0,0}^i$ on the i -th depth hypothesis plane is determined by:

$$\mathbf{c}_{n \rightarrow 0,0}^i = \mathbf{K}_0(\mathbf{R}_{n \rightarrow 0} \mathbf{K}_n^{-1} \mathbf{c}_{n,0} d_{i,0} + \mathbf{t}_{n \rightarrow 0}), \quad (2)$$

which means that:

$$\mathbf{F}_{n \rightarrow 0,0}[:, i, \mathbf{p}_{n \rightarrow 0,0}^i] = \mathbf{F}_{n,0}[:, \mathbf{p}_{n,0}], \quad (3)$$

where $[\cdot, \dots, \cdot]$ is the indexing operation of features. Similarly, the reference feature $\hat{\mathbf{F}}_{0,0}$ is also warped to the n -th source view, producing the feature $\hat{\mathbf{F}}_{0 \rightarrow n,0} \in \mathbb{R}^{C \times D_0 \times \frac{H}{8} \times \frac{W}{8}}$. During the feature projection procedure, the frustum-shaped features may lead to information loss in the spatial dimension. To mitigate this, we employ a camera embedding module composed of a multi-layer perceptron and a squeeze-and-excitation layer [11] to encode the imag-to-world camera parameters of both the reference and source views into the projected features. This process effectively injects camera parameter-related spatial information into the features and compresses the features along the depth dimension, which produces the CVPE $\mathbf{F}'_{n \rightarrow 0,0} \in \mathbb{R}^{C \times \frac{H}{8} \times \frac{W}{8}}$ and $\hat{\mathbf{F}}'_{0 \rightarrow n,0} \in \mathbb{R}^{C \times \frac{H}{8} \times \frac{W}{8}}$, as shown in Fig. 3.

The reference CVPE $\mathbf{F}'_{n \rightarrow 0,0}$ and source CVPE $\hat{\mathbf{F}}'_{0 \rightarrow n,0}$ are added to their corresponding features, which are fed into the intra-view and inter-view attention blocks. The output of the final attention block is the enhanced feature $\hat{\mathbf{F}}_{n,0}$ for the n -th view in the 0-th scale. Finally, the enhanced feature is further processed by a convolution layer and a bilinear upsampling operation to be added to the next scale of the FPN feature, producing the enhanced feature for the next scale. We denote the enhanced features of all source views as $\{\hat{\mathbf{F}}_{n,s} \in \mathbb{R}^{C \times \frac{H}{2^{3-s}} \times \frac{W}{2^{3-s}}} | s = 0, 1, 2, 3\}_{n=1}^{N-1}$.

3.3. Monocular Depth for Dynamic Depth Sampling

The monocular depth $\mathbf{D}^{mono} \in \mathbb{R}^{H \times W}$ produced by the monocular model suffers from scale ambiguity, making it unsuitable for direct application in MVS tasks. To address this, an alignment method is designed. Since the monocular model is more robust in challenging regions, we exploit the monocular depth to guide the depth sampling procedure with dynamic adjustment for edge regions.

Monocular Depth Alignment. Let $\mathbf{D}_s \in \mathbb{R}^{\frac{H}{2^{3-s}} \times \frac{W}{2^{3-s}}}$ be the predicted depth by MonoMVSNet in the s -th scale (refer to Section 3.4), we use \mathbf{D}_{s-1} to align the monocular depth \mathbf{D}^{mono} for the s -th stage. Note that in the 0-th scale, the monocular depth is directly scaled to fit the minimum and maximum depth range based on the predefined depth range.

Take the s -th ($s \geq 1$) scale for example, we first bilinearly resize the monocular depth \mathbf{D}^{mono} and the depth predicted in the previous scale \mathbf{D}_{s-1} to the resolution $(\frac{H}{2^{3-s}}, \frac{W}{2^{3-s}})$, which are respectively denoted as $\hat{\mathbf{D}}_s^{mono}$ and $\hat{\mathbf{D}}_s$. Since the depth predicted in the previous scale is coarser than that in the current scale, it is necessary to filter out the depth of unreliable pixels in $\hat{\mathbf{D}}_s$ before aligning $\hat{\mathbf{D}}_s^{mono}$ to $\hat{\mathbf{D}}_s$. To do this, we keep the coordinates of the top 80% pixels that have the largest confidence scores according to the confidence map produced in the previous scale, and the coordinates set is denoted as \mathcal{C}_s . The scale a and shift b for the alignment between $\hat{\mathbf{D}}_s^{mono}$ and $\hat{\mathbf{D}}_s$ are estimated using the least squares optimization based on these kept pixels:

$$(a, b) = \arg \min_{a, b} \sum_{\mathbf{c} \in \mathcal{C}_s} \left(\frac{1}{\hat{\mathbf{D}}_s[\mathbf{c}]} - (a\hat{\mathbf{D}}_s^{mono}[\mathbf{c}] + b)^2 \right). \quad (4)$$

The aligned monocular depth \mathbf{D} is obtained by:

$$\mathbf{D}_s^{align} = a\hat{\mathbf{D}}_s^{mono} + b. \quad (5)$$

In the following, we show how to use the aligned depth to guide the depth sampling.

Monocular Dynamic Depth Sampling. Following previous methods [31], we adopt an inverse depth sampling strategy to sampled depth candidates within the inverse depth range, satisfying the equation $\frac{1}{R_s} = \frac{1}{D_{s-1}-1} \frac{1}{R_{s-1}}$, where R_s and D_s denote the depth range and the number of depth candidates at the s -th scale, respectively.

The traditional inverse depth sampling method does not capture relative depth information, leading to blurry depth estimations in discontinuous regions, such as edges. In contrast, monocular models demonstrate impressive zero-shot performance on relative depth estimation. Intuitively, we utilize monocular depth at edge regions to guide depth sampling, ensuring more reasonable depth candidates. Specifically, we first apply a lightweight edge estimation network [28] to predict the edge confidence map $\mathbf{E} \in \mathbb{R}^{H \times W}$ for

the reference image, which is bilinearly resized to $\hat{\mathbf{E}}_s \in \mathbb{R}^{\frac{H}{2^{3-s}} \times \frac{W}{2^{3-s}}}$ for the s -th scale. Only the pixels that have the edge confidence larger than a pre-defined threshold λ are kept, and their coordinates set is denoted as \mathcal{C}_s^{edge} .

Supposing the depth hypotheses for the s -th scale are $\{d_{i,s}\}_{i=0}^{D_s-1}$ during the inverse depth sampling procedure. For each pixel with the coordinate $\mathbf{c} \in \mathcal{C}_s^{edge}$, the absolute difference between the depth value $\mathbf{D}_s^{align}[\mathbf{c}]$ to all D_s depth candidate values are computed. The depth candidate value that has the smallest absolute difference is replaced by the value $\mathbf{D}_s^{align}[\mathbf{c}]$. The remaining pixels whose coordinates are not in \mathcal{C}_s^{edge} share the same depth candidates $\{d_{i,s}\}_{i=0}^{D_s-1}$. As we can see, the depth candidates are dynamically updated for the edge regions, which captures accurate relative depth information from the monocular depth.

3.4. Depth Prediction with Cost Volume

Following existing works [10, 35], the depth \mathbf{D}_s in the s -th scale is obtained with a cost volume, which is constructed based on the correlations between the reference feature and all the warped source features. The overall procedure can be divided into 3 steps: (1) warping all source features from the source view to the reference view, (2) constructing cost volume with feature correlations, and (3) predicting the depth based on the cost volume. For more details about the construction of cost volume, please refer to the works in [31, 50]. Let $\mathbf{V}_s \in \mathbb{R}^{G \times D_s \times \frac{H}{2^{3-s}} \times \frac{W}{2^{3-s}}}$ be the constructed cost value, it is fed into a lightweight 3D UNet regularization network [31] (followed by a softmax function) to obtain the probability volume $\mathbf{P}_s \in \mathbb{R}^{D_s \times \frac{H}{2^{3-s}} \times \frac{W}{2^{3-s}}}$. The probability volume \mathbf{P}_s is then used to predict the depth \mathbf{D}_s through a winners-take-all strategy. Take the pixel with coordinate \mathbf{c} for example:

$$\mathbf{D}_s[\mathbf{c}] = d_{i,s} \quad \text{s.t.} \quad \arg \max_i \mathbf{P}[i, \mathbf{c}]. \quad (6)$$

3.5. Training Objective with Relative Consistency

To further leverage the outstanding relative depth estimation capability of the monocular model, we propose a relative consistency loss to supervise the the depth produced by MonoMVSNet. Take the s -th scale for example, we first get the depth with probability $\mathbf{D}_s^{prob} \in \mathbb{R}^{\frac{H}{2^{3-s}} \times \frac{W}{2^{3-s}}}$. Take the pixel with coordinate \mathbf{c} for example:

$$\mathbf{D}_s^{prob}[\mathbf{c}] = \sum_{i=0}^{D_s} \mathbf{P}[i, \mathbf{c}] d_{i,s}. \quad (7)$$

Then we randomly sample two sets of pixels with each set contains M pixels. The corresponding coordinate sets are denoted as \mathcal{C}_s^1 and \mathcal{C}_s^2 . The relative consistency loss is:

$$\mathcal{L}_s^{rc} = \frac{1}{M} \sum_{m=0}^{M-1} \max(0, -e_m), \quad (8)$$

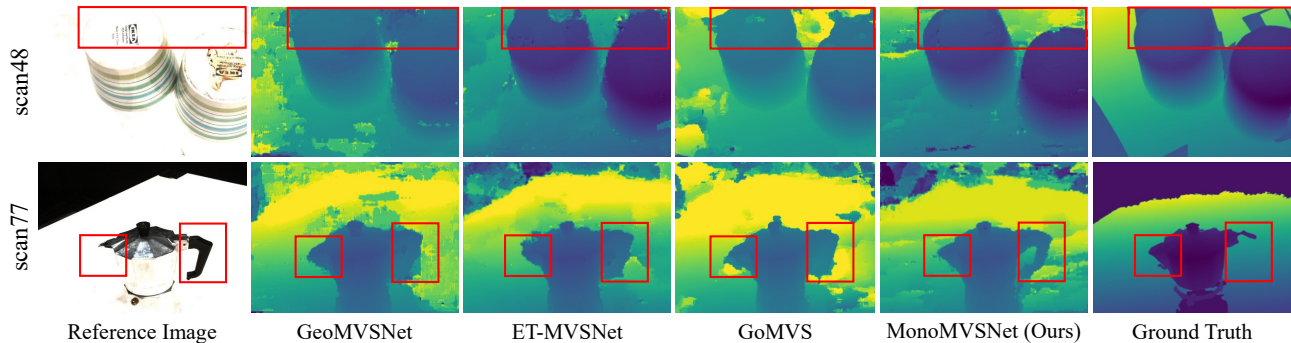


Figure 4. Qualitative comparison of predicted depth maps with GeoMVSNet [50], ET-MVSNet [20] and GoMVSNet [33] on the most challenge scenes, scan48 and scan77 from DTU test set. Our method predicts more accurate depth in reflective surfaces and depth-discontinuous edge regions.

where e_m is the relative error between the m -th pixel in these two sets:

$$e_m = (\mathbf{D}_s^{prob}[\mathbf{c}_m^1] - \mathbf{D}_s^{prob}[\mathbf{c}_m^2]) \cdot \text{Sign}(\hat{\mathbf{D}}_s^{mono}[\mathbf{c}_m^1] - \hat{\mathbf{D}}_s^{mono}[\mathbf{c}_m^2]), \quad (9)$$

where $\mathbf{c}_m^1 \in \mathcal{C}_s^1$ and $\mathbf{c}_m^2 \in \mathcal{C}_s^2$ are the coordinates of the m -th pixel in the two sets, and $\text{Sign}(\cdot)$ is the function that returns the sign (*i.e.*, -1 or +1) of the given value.

The overall training objective is composed of two primary components. First, there’s the standard cross-entropy loss \mathcal{L}_s^{ce} , which supervises the predicted probability volume \mathbf{P}_s against the ground truth volume for each scale s . Second, we introduce the relative consistency loss \mathcal{L}_s^{rc} , a novel component exclusively applied at the final scale. Formally, the overall training objective of MonoMVSNet is:

$$\mathcal{L}_{overall} = \sum_{s=0}^3 \mathcal{L}_s^{ce} + \gamma \mathcal{L}_3^{rc}. \quad (10)$$

4. Experiments

4.1. Datasets and Metrics

Dataset. DTU [1] is a object-centered indoor dataset that includes 128 scenes, with each scene containing 49 or 64 images captured under 7 different lighting conditions. Following the standard practice [41], the dataset is divided into training, test, and validation sets. Tanks-and-Temples [16] dataset is a large-scale outdoor real-world scene dataset with complex transformations and lighting variations, divided into an intermediate subset with 8 scenes and an advanced subset with 6 scenes. BlendedMVS [43] dataset is a large-scale synthetic dataset that includes both indoor and outdoor scenes, offering a training set with 106 scenes and a validation set with 7 scenes.

Metrics. To validate the effectiveness of the proposed MonoMVSNet, we compute metrics for both point clouds

Methods	Years	Overall↓	Acc.↓	Comp.↓
Gipuma [8]	ICCV’15	0.578	0.283	0.873
COLMAP [27]	CVPR’16	0.532	0.400	0.664
MVSNet [41]	ECCV’18	0.462	0.396	0.527
CasMVSNet [9]	CVPR’20	0.355	0.325	0.385
UniMVSNet [24]	CVPR’22	0.315	0.352	0.278
TransMVSNet [7]	CVPR’22	0.305	0.321	0.289
MVSTER* [31]	ECCV’22	0.313	0.350	0.276
WT-MVSNet [18]	NIPS’22	0.295	0.309	0.281
RA-MVSNet [49]	CVPR’23	0.297	0.326	0.268
GeoMVSNet [50]	CVPR’23	0.295	0.331	0.259
DMVSNet [44]	ICCV’23	0.305	0.338	0.272
ET-MVSNet [20]	ICCV’23	0.291	0.329	0.253
MVSFormer* [2]	TMLR’23	0.289	0.327	0.251
DS-PMNet [17]	AAAI’24	0.290	0.323	0.257
MVSFormer++* [3]	ICLR’24	0.281	0.309	0.252
GoMVS [33]	CVPR’24	0.287	0.347	0.227
MonoMVSNet (Ours)	$N=5$	<u>0.278</u>	0.313	<u>0.243</u>
MonoMVSNet (Ours)	$N=9$	0.275	<u>0.302</u>	0.248

Table 1. Quantitative results of reconstructed point clouds on the DTU [1] evaluation set with distance metrics [mm]. The best and second best values are highlighted with **bold** and underline. Methods with * denotes using high-resolution images for training.

and depth. For the point cloud metric, we use the official MATLAB code and evaluation dataset provided by DTU [1] to measure the distance between the generated point clouds and the ground truth, reporting accuracy (Acc.), completeness (Comp.), and their average, Overall. Additionally, for the Tanks-and-Temples [16] dataset, we evaluate the generated point clouds by uploading them to the official website, reporting F-score in percentage. For the depth metric, we report the Mean Absolute Error (MAE) and depth error ratios at 2mm (e_2), 4mm (e_4), and 8mm (e_8) on DTU.

4.2. Implementation Details

MonoMVSNet was developed using PyTorch [23] and utilizes the Adam [15] optimizer. For the DTU [1] dataset, the model is trained for 15 epochs using 5-view input images at

Methods	Years	Intermediate subset†									Advanced subset†						
		Mean	Fam.	Fran.	Hor.	Lig.	M60	Pan.	Pla.	Tra.	Mean	Aud.	Bal.	Cou.	Mus.	Pal.	Tem.
COLMAP [27]	CVPR'16	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 [9]	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
TransMVSNet [7]	CVPR'22	63.52	80.92	65.83	56.94	62.54	63.06	60.00	60.20	58.67	37.00	24.84	44.59	34.77	46.49	34.69	36.62
UniMVSNet [24]	CVPR'22	64.36	81.20	66.43	53.11	63.46	66.09	64.84	62.23	57.53	38.96	28.33	44.36	39.74	52.89	33.80	34.63
MVSTER [31]	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
WT-MVSNet [18]	NIPS'22	65.34	81.87	67.33	57.76	64.77	65.68	64.61	62.35	58.38	39.91	29.20	44.48	39.55	53.49	34.57	38.15
CostFormer [2]	IJCAI'23	64.51	81.31	65.65	55.57	63.46	<u>66.24</u>	65.39	61.27	57.30	39.43	29.18	45.21	39.88	53.38	34.07	34.87
RA-MVSNet [49]	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
GeoMVSNet [50]	CVPR'23	65.89	81.64	67.53	55.78	68.02	65.49	<u>67.19</u>	<u>63.27</u>	58.22	41.52	30.23	46.54	39.98	53.05	35.98	43.34
DMVSNet [44]	ICCV'23	64.66	81.27	67.54	59.10	63.12	64.64	64.80	59.83	56.97	41.17	30.08	46.10	40.65	53.53	35.08	41.60
ET-MVSNet [20]	ICCV'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
MVSFormer [2]	TMLR'23	66.37	82.06	69.34	60.49	68.61	65.67	64.08	61.23	59.53	40.87	28.22	46.75	39.30	52.88	35.16	42.95
DS-PMNet [17]	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
MVSFormer++ [3]	ICLR'24	67.18	82.69	<u>69.44</u>	<u>64.24</u>	<u>69.16</u>	64.13	66.43	61.19	60.12	41.60	29.93	45.69	39.46	<u>53.58</u>	35.56	<u>45.39</u>
GoMVS [33]	CVPR'24	66.44	<u>82.68</u>	69.23	69.19	63.56	65.13	62.10	58.81	60.80	<u>43.07</u>	35.52	47.15	<u>42.52</u>	52.08	<u>36.34</u>	44.82
MonoMVSNet (Ours)	-	68.63	82.38	72.89	62.80	70.49	65.79	68.54	65.54	<u>60.59</u>	43.58	<u>30.33</u>	<u>46.76</u>	42.90	56.31	37.28	47.88

Table 2. Quantitative results on the Tanks-and-Temples benchmark with F-score [%]. **Bold** figures represent the best and underline figures represent the second best, respectively. The mean refers the average F-score of all scenes.

a resolution of 512×640 with a batch size of 4. The initial learning rate is 0.001, which is halved after the 10-th, 12-th, and 14-th epochs. For the BlendedMVS dataset, we fine-tune the model trained on DTU for 15 epochs using 9-view images at a resolution of 576×768 with a batch size of 4. The initial learning rate is 0.001, which is halved after the 6-th, 8-th, 10-th, and 12-th epochs. We performe inverse depth sampling [35] within a depth range from $425mm$ to $935mm$, with depth hypotheses in the coarse-to-fine stages as 8-8-4-4, depth intervals of 0.5-0.5-0.5-0.5, and group correlations of 8-8-4-4.

4.3. Benchmark Performance

Evaluation on DTU. We compare MonoMVSNet with traditional and learning-based methods. We predict depth using 5-view images with a resolution of 832×1152 and employ a dynamic fusion strategy [37] for point cloud reconstruction. The quantitative results of the point clouds are shown in Tab. 1, where MonoMVSNet achieves the highest overall performance among all methods and ranks second in accuracy, only behind the traditional method Gipuma [8], which performs poorly overall performance. Notably, MonoMVSNet’s memory consumption during inference is 2.01GB, which is significantly better than recent state-of-the-art methods [2, 3, 33], as illustrated in Fig. 1 row 1 (a). The qualitative comparison of depth maps is presented in Fig. 4, where our method predicts more accurate depth maps in the most challenging scenes, scan48 and scan77.

Evaluation on Tanks-and-Temples. We further evaluate the generalization capability of MonoMVSNet on the Tanks-and-Temples [16] benchmark. Consistent with [2, 3], we set the number of input images $N = 21$ with 2k resolution. A dynamic fusion strategy is adopted for point cloud reconstruction, and the quantitative comparison results are presented in Tab. 2. Our method achieves the highest F-score on both the intermediate and advanced subsets, demonstrating its strong generalization ability.

4.4. Ablation Study

Ablation study is conducted on DTU [1] to verify the effectiveness of each components. Unless otherwise specified, we use 5-view input images with a resolution of 832×1152 and the dynamic fusion strategy [37] for point cloud reconstruction, with all other hyperparameters kept consistent. The overall results of the ablation are presented in Tab. 3. Our method achieves significant improvements in both point cloud and depth metrics, demonstrating the effectiveness of our proposed method.

Effectiveness of Monocular Feature. Compared to the baseline model, introducing monocular features improves the overall point cloud metric from 0.303 to 0.288, and increased the depth metric MAE by 12.4%. As shown in Tab. 3 model (a)(b)(c), using reference monocular features (RMF) improves point cloud completeness from 0.282 to 0.263, indicating that monocular features effectively address areas where matching fails. Additionally, existing cross attention (CA) slightly improves the depth metrics, primarily because traditional positional encoding is not specifically designed for MVS tasks and lacks an understanding of three-dimensional spatial context. However, the introduction of CVPE significantly enhances the performance in both point cloud and depth metrics.

Effectiveness of Monocular Depth. The exploitation of monocular depth boosts the overall performance score from 0.292 to 0.281. As shown in Tab. 3, monocular dynamic depth sampling (MDDS) only yields improvements in depth metrics. However, with the introduction of the edge map (EM) shows improvements in both point cloud and depth performance, indicating that excessive monocular sampling could introduce erroneous depth candidates. The relative consistency loss (RCL) balances optimization by enforcing consistency between multi-view depth and monocular depth, further enhancing model performance.

Feature Extraction Design. As shown in Tab. 4, we inves-

Models	Monocular Feature			Monocular Depth			Overall↓	Acc.↓	Comp.↓	MAE↓	e_2 ↓	e_4 ↓	e_8 ↓
	RMF	CA	CVPE	MDDS	EM	RCL							
base							0.303	0.323	0.282	6.53	20.49	12.85	8.61
(a)	✓						0.293	0.322	0.263	5.86	19.08	11.61	7.55
(b)	✓	✓					0.293	0.324	0.262	5.82	19.41	11.76	7.61
(c)	✓	✓	✓				0.288	0.317	0.259	5.72	18.69	11.33	7.37
(d)	✓	✓	✓	✓			0.292	0.318	0.265	5.44	17.55	10.60	6.82
(e)	✓	✓	✓	✓	✓		0.283	0.305	0.260	5.04	16.60	9.81	6.31
full	✓	✓	✓	✓	✓	✓	0.281	0.314	0.248	4.99	16.40	9.75	6.27

Table 3. Ablation study on DTU evaluation set, using normal fusion strategy for point cloud reconstruction. We analyze the effects of the reference monocular feature (RMF), cross attention (CA), cross-view positional encoding (CVPE), monocular guided depth sampling (MDDS), edge map (EM), and relative consistency loss (RCL).

Feature Encoder	Overall↓	MAE↓	Memory↓	Time↓
All (w/o CVPE)	0.297	6.13	5.72GB	0.55s
All (w/ CVPE)	0.296	5.82	5.72GB	0.56s
Ref. (w/o CVPE)	0.284	5.37	2.01GB	0.24s
Ref. (w/ CVPE)	0.278	4.99	2.01GB	0.25s

Table 4. Ablation study on feature extraction design.

tigate the performance of different feature extraction strategies. “All” refers to extracting monocular features from all input images, while “Ref.” indicates extracting monocular features only from the reference image. As illustrated in rows 1 and 3, “Ref.” outperforms “All” in both point cloud and depth metrics, and is more efficient (GPU memory reduced by 64.9%, run time reduced by 56.4%). We speculate this is due to the performance gap between different models—excessive monocular features might negatively affect performance. Additionally, the proposed CVPE design further boosts the performance of the “Ref.” model with negligible computational cost (rows 3 and 4).

ViT Variants. We conduct the ablation study on different ViT variants of monocular foundation models, as shown in Tab. 5. The ViT-S and ViT-B variants exhibit similar performance, while the ViT-L variant shows a decrease in performance. This indicates that the small (ViT-S) variant of monocular foundation models can provide sufficient prior information without relying on highly parameterized base (ViT-B) or large (ViT-L) models.

Model Efficiency. We compare MonoMVSNet with several state-of-the-art methods regarding runtime, GPU memory consumption, and parameter count. The results are summarized in Table 6. MonoMVSNet consumes the least GPU memory compared to methods without pre-trained models (GeoMVSNet [50], ET-MVSNet [20], and GoMVS [33]) and methods based on pre-trained models (MVSFormer [2], MVSFormer++ [3]). This advantage stems from its uniquely designed feature extraction strategy. Additionally, MonoMVSNet exhibits significantly faster runtime

Backbones	Overall↓	Acc.↓	Comp.↓	MAE↓
ViT-S	0.278	0.313	0.243	4.99
ViT-B	0.278	0.294	0.262	4.98
ViT-L	0.281	0.311	0.251	4.94

Table 5. Ablation study on different ViT variants.

Methods	Memory	Time	Params(all)	Params(train)
GeoMVSNet [50]	5.21GB	0.19s	15.31M	15.31M
ET-MVSNet [20]	2.91GB	0.16s	1.09M	1.09M
GoMVS [33]	12.61GB	0.64s	1.50M	1.50M
MVSFormer [2]	3.66GB	0.24s	28.01M	28.01M
MVSFormer++ [3]	4.71GB	0.23s	126.95M	39.48M
MonoMVSNet	2.01GB	0.25s	27.68M	2.89M

Table 6. Memory, time and parameters comparison per image at a resolution of 832×1152 with 5-view images.

than GoMVS and possesses substantially fewer trainable parameters than similar models (MVSFormer and MVSFormer++). These results clearly highlight the superior efficiency of our proposed MonoMVSNet.

5. Conclusion

In this paper, we present a monocular-priors-guided multi-view stereo network, MonoMVSNet. The monocular feature and monocular depth from the pre-trained monocular foundation model are efficiently and elegantly integrated into feature extraction and depth sampling procedures. In addition, a relative consistency loss is designed to supervise the prediction depth with the monocular depth. With the help of the priors in pre-trained monocular model, MonoMVSNet can predict more accurate depth, especially for the depth-discontinuous regions (*e.g.*, edge regions). Experimental results demonstrate that our method outperforms state-of-the-art methods on different datasets with less GPU and inference time consumption.

Acknowledgments. This work was supported by the Beijing Natural Science Foundation (No. L257003), National Natural Science Foundation of China (No. 62402042 and 62227801).

References

- [1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjarholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120:153–168, 2016. 1, 6, 7
- [2] Chenjie Cao, Xinlin Ren, and Yanwei Fu. Mvsformer: Multi-view stereo by learning robust image features and temperature-based depth. *Transactions on Machine Learning Research*, 2022. 1, 2, 6, 7, 8
- [3] Chenjie Cao, Xinlin Ren, and Yanwei Fu. Mvsformer++: Revealing the devil in transformer’s details for multi-view stereo. *arXiv preprint arXiv:2401.11673*, 2024. 1, 2, 6, 7, 8
- [4] Shuo Cheng, Zexiang Xu, Shilin Zhu, Zhuwen Li, Li Erran Li, Ravi Ramamoorthi, and Hao Su. Deep stereo using adaptive thin volume representation with uncertainty awareness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2524–2534, 2020. 2
- [5] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. *Advances in neural information processing systems*, 34:9355–9366, 2021. 2
- [6] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 764–773, 2017. 2
- [7] 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. 1, 2, 4, 6, 7
- [8] 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. 1, 6, 7
- [9] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504, 2020. 1, 2, 6, 7
- [10] 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. 5
- [11] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 4
- [12] Jianfei Jiang, Mingwei Cao, Jun Yi, and Chenglong Li. Di-mvs: Learning efficient multi-view stereo with depth-aware iterations. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3180–3184, 2024. 1
- [13] Jianfei Jiang, Liyong Wang, Haochen Yu, Tianyu Hu, Jiansheng Chen, and Huimin Ma. Rrt-mvs: Recurrent regularization transformer for multi-view stereo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3994–4002, 2025. 1
- [14] Seung-Wook Kim, Hyong-Keun Kook, Jee-Young Sun, Mun-Cheon Kang, and Sung-Jea Ko. Parallel feature pyramid network for object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 234–250, 2018. 1, 2
- [15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6
- [16] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017. 1, 6, 7
- [17] Hongjie Li, Yao Guo, Xianwei Zheng, and Hanjiang Xiong. Learning deformable hypothesis sampling for accurate patchmatch multi-view stereo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3082–3090, 2024. 6, 7
- [18] Jinli Liao, Yikang Ding, Yoli Shavit, Dihe Huang, Shihao Ren, Jia Guo, Wensen Feng, and Kai Zhang. Wt-mvsnet: window-based transformers for multi-view stereo. *Advances in Neural Information Processing Systems*, 35:8564–8576, 2022. 2, 4, 6, 7
- [19] Hongyuan Liu, Haochen Yu, Bochao Zou, Juntao Lyu, Qi Mei, Jiansheng Chen, and Huimin Ma. Protocar: Learning 3d vehicle prototypes from single-view and unconstrained driving scene images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5460–5468, 2025. 1
- [20] Tianqi Liu, Xinyi Ye, Weiyue Zhao, Zhiyu Pan, Min Shi, and Zhiguo Cao. When epipolar constraint meets non-local operators in multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18088–18097, 2023. 2, 4, 6, 7, 8
- [21] Zhenxing Mi, Chang Di, and Dan Xu. Generalized binary search network for highly-efficient multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12991–13000, 2022. 2, 3
- [22] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2
- [23] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6
- [24] Rui Peng, Rongjie Wang, Zhenyu Wang, Yawen Lai, and Ronggang Wang. Rethinking depth estimation for multi-view stereo: A unified representation. In *Proceedings of*

- the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8645–8654, 2022. 6, 7
- [25] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 3
- [26] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 3
- [27] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part III 14*, pages 501–518. Springer, 2016. 6, 7
- [28] Xavier Soria, Yachuan Li, Mohammad Rouhani, and Angel D Sappa. Tiny and efficient model for the edge detection generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1364–1373, 2023. 5
- [29] 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. 2
- [30] Likang Wang, Yue Gong, Xinjun Ma, Qirui Wang, Kaixuan Zhou, and Lei Chen. Is-mvsnet: importance sampling-based mvsnet. In *European Conference on Computer Vision*, pages 668–683. Springer, 2022. 3
- [31] 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. 5, 6, 7
- [32] 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. 2
- [33] Jiang Wu, Rui Li, Haofei Xu, Wenxun Zhao, Yu Zhu, Jinqiu Sun, and Yanning Zhang. Gomvs: Geometrically consistent cost aggregation for multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20207–20216, 2024. 6, 7, 8
- [34] Qingshan Xu and Wenbing Tao. Planar prior assisted patch-match multi-view stereo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12516–12523, 2020. 1
- [35] 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*, pages 12508–12515, 2020. 5, 7
- [36] Qingshan Xu, Weihang Kong, Wenbing Tao, and Marc Pollefeys. Multi-scale geometric consistency guided and planar prior assisted multi-view stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4945–4963, 2022. 1
- [37] Jianfeng Yan, Zizhuang Wei, Hongwei Yi, Mingyu Ding, Runze Zhang, Yisong Chen, Guoping Wang, and Yu-Wing Tai. Dense hybrid recurrent multi-view stereo net with dynamic consistency checking. In *European conference on computer vision*, pages 674–689. Springer, 2020. 2, 7
- [38] Jiayu Yang, Wei Mao, Jose M Alvarez, and Miaomiao Liu. Cost volume pyramid based depth inference for multi-view stereo. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4877–4886, 2020. 2
- [39] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 2, 3
- [40] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. 1, 2, 3, 4
- [41] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pages 767–783, 2018. 1, 2, 6
- [42] 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. 1, 2
- [43] 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. 6
- [44] Xinyi Ye, Weiyue Zhao, Tianqi Liu, Zihao Huang, Zhiguo Cao, and Xin Li. Constraining depth map geometry for multi-view stereo: A dual-depth approach with saddle-shaped depth cells. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17661–17670, 2023. 6, 7
- [45] Haochen Yu, Weixi Gong, Jiansheng Chen, and Huimin Ma. Get3dgs: Generate 3d gaussians based on points deformation fields. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 1
- [46] Zhenlong Yuan, Cong Liu, Fei Shen, Zhaoxin Li, Jinguo Luo, Tianlu Mao, and Zhaoqi Wang. Msp-mvs: Multi-granularity segmentation prior guided multi-view stereo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9753–9762, 2025. 1
- [47] Zhenlong Yuan, Jinguo Luo, Fei Shen, Zhaoxin Li, Cong Liu, Tianlu Mao, and Zhaoqi Wang. Dvp-mvs: Synergize depth-edge and visibility prior for multi-view stereo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9743–9752, 2025.
- [48] Zhenlong Yuan, Zhidong Yang, Yujun Cai, Kuangxin Wu, Mufan Liu, Dapeng Zhang, Hao Jiang, Zhaoxin Li, and Zhaoqi Wang. Sed-mvs: Segmentation-driven and edge-aligned deformation multi-view stereo with depth restoration

- and occlusion constraint. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 1
- [49] Yisu Zhang, Jianke Zhu, and Lixiang Lin. Multi-view stereo representation revisited: Region-aware mvsnet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17376–17385, 2023. 6, 7
- [50] Zhe Zhang, Rui Peng, Yuxi Hu, and Ronggang Wang. Geomvsnet: Learning multi-view stereo with geometry perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21508–21518, 2023. 5, 6, 7, 8
- [51] Jihuai Zhao, Junbao Zhuo, Jiansheng Chen, and Huimin Ma. Sam2object: Consolidating view consistency via sam2 for zero-shot 3d instance segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19325–19334, 2025. 1