BridgeMVS: Bridging Multi-View Stereo and Monodepth via Bidirectional Dynamic Fusion

Jianfei Jiang* **Haochen Yu*** **Qiankun Liu†** **Rui Yang**
Hongyuan Liu **Xianglong Meng** **Jiansheng Chen** **Huimin Ma†**
University of Science and Technology Beijing, China
{jiangjf,haochen.yu,ruiyang,hongyuanliu,mengxianglong}@xs.ustb.edu.cn
{liuqk3,jschen,mhmpub}@ustb.edu.cn

Abstract

Multi-View Stereo (MVS) and Monocular Depth Estimation (MDE) provide complementary cues for dense 3D reconstruction. While MVS leverages explicit geometry for metric depth, it struggles in challenging regions where feature matching is unreliable. Conversely, MDE provides robust structural priors but typically yields scale-ambiguous results. Existing hybrid methods mainly exploit monocular cues in a unidirectional manner, leaving the mutual enhancement between MDE and MVS insufficiently explored. To address this limitation, we propose BridgeMVS, a unified cascaded multi-stage MVS framework that bridges monocular depth features and multi-view cost volume representations through Bidirectional Dynamic Fusion (BDF). At each stage, BDF performs dynamic mono-to-cost and cost-to-mono updates, where monocular structural features guide cost volume refinement for more robust MVS depth estimation, while multi-view geometric evidence recalibrates monocular representations to improve their metric-scale consistency. Furthermore, the fused monocular and multi-view representations are progressively propagated from coarse to fine stages and constrained by stage-wise relative-scale MDE supervision and metric-scale MVS supervision, enabling cross-branch interactions to benefit subsequent high-resolution depth estimation. In addition, BDF is designed as a plug-and-play module that can be conveniently integrated into existing cascaded MVS frameworks. Extensive experiments demonstrate that BridgeMVS achieves state-of-the-art performance on the DTU, Tanks-and-Temples, and ETH3D benchmarks, while simultaneously boosting the accuracy of monocular depth estimation. Notably, BridgeMVS ranks first on both the Tanks-and-Temples Intermediate and Advanced benchmarks. The source code is available at https://github.com/JianfeiJ/BridgeMVS.

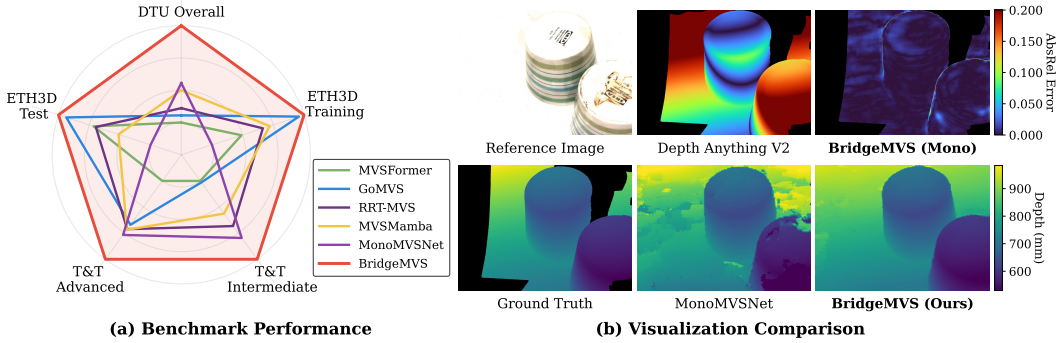
1 Introduction

Multi-View Stereo (MVS) is a fundamental task in computer vision that aims to reconstruct dense 3D geometry from a set of calibrated images, which plays a crucial role in various applications [1, 2, 3, 4, 5]. Traditional MVS methods [6, 7, 8, 9, 10, 11] rely on handcrafted photo-consistency metrics, which limits their robustness in real-world scenarios. Learning-based MVS methods [12, 13] replace handcrafted matching costs with learned feature representations and differentiable cost volumes constructed via the plane-sweep algorithm [14], achieving substantial progress.

Despite this progress, existing learning-based MVS methods still struggle in challenging regions, such as weakly textured areas and reflective surfaces. In these regions, reliable correspondences

*Equal contribution.
†Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Benchmark Performance **(b) Visualization Comparison**

Figure 1: (a) Benchmark performance. Our method outperforms state-of-the-art methods [33, 34, 32, 20, 35] on three standard MVS benchmarks [36, 37, 38]. (b) Visualization comparison on Scan 48 from DTU [36]. Benefiting from bidirectional dynamic fusion, BridgeMVS improves MVS depth estimation [35] in challenging regions and enhances the accuracy of monocular depth estimation [39] under multi-view geometric constraints. Absolute relative error maps are visualized after aligning the relative depth with GT using the ROE solver [40].

across views are difficult to establish, resulting in inaccurate matching costs and ambiguous cost volumes. Existing methods mainly address this issue from two perspectives. One line of work improves feature representation by introducing CNN variants [15], Transformers [16, 17, 18], and Mamba [19, 20] for feature aggregation. Another line focuses on cost volume regularization, evolving from single-stage 3D CNNs and RNNs [21, 22, 23, 24, 25] to cascade frameworks [26, 27, 28], and further incorporating Transformers [29, 30, 31, 18, 32] to aggregate matching costs and suppress unreliable responses. Nevertheless, the cost volume is still fundamentally built upon multi-view feature matching, making MVS vulnerable when visual correspondences are unreliable.

Recent Monocular Depth Estimation (MDE) models [41, 39] provide promising complementary cues for MVS by encoding rich structural priors from large-scale training. These priors are naturally complementary to MVS: MDE offers strong structural awareness but predicts scale-ambiguous relative depth, while MVS provides explicit multi-view geometric constraints and metric-scale depth. Recent MonoMVSNet [35] explores this complementarity by introducing monocular priors into the MVS pipeline. However, it mainly follows a unidirectional paradigm, where MDE only serves as external guidance for MVS. Such one-way interaction limits the use of monocular priors and prevents MDE from benefiting from multi-view geometric evidence, making it difficult to jointly improve metric MVS depth and relative monocular depth, as shown in Fig. 1.

To address these limitations, we propose BridgeMVS, a unified cascaded MVS framework that bridges monocular depth features and multi-view cost volume representations through Bidirectional Dynamic Fusion (BDF). Instead of simply injecting monocular cues into the MVS branch, BDF performs bidirectional feature interaction at each cascade stage. In the mono-to-cost direction, monocular structural features are adapted to the cost volume space and used to guide cost volume refinement, helping the MVS branch distinguish ambiguous or redundant matching responses. In the cost-to-mono direction, geometry-aware information is aggregated from the regularized cost volume and fed back to the MDE branch, recalibrating intermediate monocular representations with explicit multi-view evidence. Through this bidirectional update, BridgeMVS jointly improves multi-view reconstruction and monocular depth estimation, leading to superior benchmark performance and more accurate depth predictions, as illustrated in Fig. 1.

Furthermore, BridgeMVS exploits the inherent multi-scale representations of the DPT-based MDE branch and aligns them with the cascaded MVS structure. The fused monocular and multi-view representations are progressively propagated from coarse to fine stages, allowing early cross-branch interactions to benefit subsequent high-resolution depth estimation. Meanwhile, stage-wise relative-scale MDE supervision and metric-scale MVS supervision provide complementary structural and geometric constraints throughout the cascade. By preserving fused representations across stages, BridgeMVS enables monocular priors and multi-view geometric evidence to be gradually refined. Since BDF operates only on intermediate monocular features and cost volume representations, it can be used as a plug-and-play module for existing cascaded MVS frameworks.

The main contributions of this work are summarized as follows:

- We propose a plug-and-play Bidirectional Dynamic Fusion (BDF) module that performs mono-to-cost and cost-to-mono updates between monocular depth features and multi-view cost volumes, and can be conveniently integrated into existing cascaded MVS frameworks.
- We propagate fused monocular and multi-view representations across cascaded stages from coarse to fine, and introduce stage-wise relative-scale and metric-scale supervision to enable progressive multi-scale refinement under both structural and geometric constraints.
- Extensive experiments show that BridgeMVS achieves state-of-the-art performance on multiple MVS benchmarks, while also substantially improving monocular depth estimation.

2 Related Work

2.1 Learning-based MVS

Multi-View Stereo (MVS) reconstructs 3D geometry from calibrated multi-view images. MVS-Net [12] firstly introduces a plane-sweep cost volume with 3D CNN-based regularization, establishing the standard learning-based MVS pipeline of feature extraction, cost volume construction, cost regularization, and depth prediction. Recent methods mainly improve MVS from two aspects. The first focuses on feature representation. Transformer-based methods [16, 17, 33, 18] enhance intra-view and inter-view feature aggregation with attention mechanisms, while MVSMamba [20] introduces state-space modeling for efficient global feature aggregation. The second focuses on cost volume regularization. WT-MVSNet [30], CostFormer [31], and RRT-MVS [32] adopt window attention [42] to aggregate matching costs, while MVSFormer++ [18] uses flash attention [43] for cost regularization. GoMVS [34] further incorporates normal priors to address local geometric inconsistency. Despite these advances, the cost volume is still fundamentally constrained to be constructed from multi-view feature matching, making MVS unreliable in weakly textured, reflective, or repetitive regions.

2.2 MVS with MDE

Monocular Depth Estimation (MDE) [44] predicts depth from a single image and provides rich contextual and structural priors. Early models such as MiDaS [45] and DPT [46] improve cross-dataset generalization with robust dense prediction architectures. Recent monocular foundation models, including Depth Anything [41, 39] and MoGe [40, 47], further demonstrate strong zero-shot generalization and produce coherent depth structures in challenging regions. Monocular priors have recently been introduced into stereo matching [48, 49, 50, 51, 52] and MVS to improve robustness when correspondence estimation is ambiguous. Among stereo matching methods, MonSter [49] performs bidirectional refinement in the disparity space through scale-shift alignment and recurrent updates, while BridgeDepth [50] employs cross-attention for single-stage feature alignment after DPT fusion. For MVS, MVSAnywhere [53] proposes a cost volume patchifier module to fuse cost volume and monocular encoder feature, and leverages large-scale training for zero-shot MVS depth estimation. MonoMVSNet [35] introduces monocular features and monocular depth priors through feature guidance, dynamic depth sampling, and relative-consistency supervision. Concurrent work SPE-MVS [54] incorporates monocular depth priors through spatial position encoding to enhance reconstruction performance. These MVS methods demonstrate the effectiveness of monocular priors but primarily use them to guide MVS in a unidirectional manner. In contrast, our method establishes multi-stage bidirectional interaction between intermediate DPT features and regularized multi-view cost volumes, enabling mutual enhancement of both MVS and MDE.

3 Method

3.1 Overview

The architecture of BridgeMVS is shown in Fig. 2. Given N calibrated images, including one reference image I_0 and $N-1$ source images $\{I_i\}_{i=1}^{N-1}$, BridgeMVS estimates both a metric MVS depth map and a relative monocular depth map for the reference image within a cascaded multi-stage

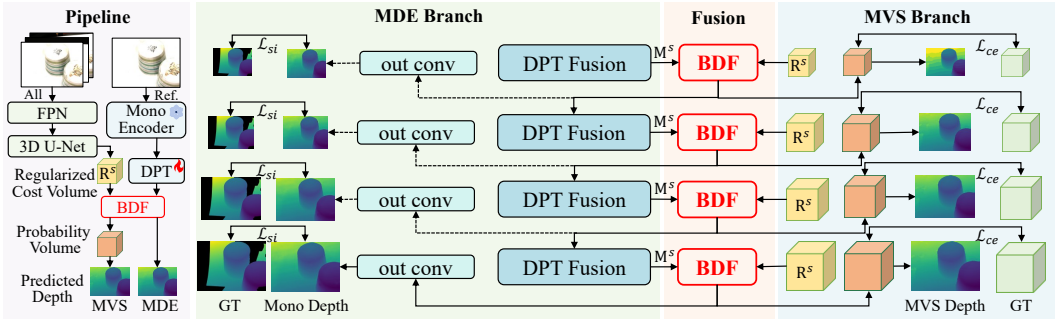


Figure 2: Overall architecture of BridgeMVS. The framework contains a cost volume-based MVS branch, a DPT-based MDE branch, and a Bidirectional Dynamic Fusion (BDF) module at each cascade stage. BDF bidirectionally fuses regularized cost volumes and monocular features, producing enhanced multi-view and monocular representations that are progressively propagated from coarse to fine stages under metric-scale MVS supervision and relative-scale MDE supervision.

MVS framework. The network consists of a cost volume-based MVS branch, a DPT-based MDE branch [39], and a Bidirectional Dynamic Fusion (BDF) module (Sec. 3.2) inserted at each stage.

MVS Branch follows the main pipeline of existing cascaded multi-scale MVS methods [26, 55] except for the inserted BDF module. Specifically, all images are fed into a Feature Pyramid Network (FPN) [56] to extract multi-scale image features. At stage s, each source-view feature is warped to the reference view to compute group-wise feature correlations [57, 58] with the reference-view feature over the predefined depth hypotheses. Pixel-view weights [59] are then used to fuse all pairwise correlations into a single cost volume. Following common practice, a lightweight 3D U-Net [55] is applied for cost volume regularization. Instead of directly predicting the metric MVS depth from the regularized cost volume $\mathbf{R}^s$, BridgeMVS couples the regularized cost volume with the monocular feature (from the corresponding stage of the MDE branch) through BDF. The BDF-refined cost volume is converted into a probability volume by applying softmax along the depth dimension, and the metric MVS depth is obtained using a winner-take-all strategy. More details on cascaded multi-scale MVS can be found in [26, 55].

MDE Branch follows the pipeline of the monocular depth foundation model (e.g., Depth Anything V2 [39]) except for the BDF module. It consists of a DINOv2-based encoder [60] and a DPT-based decoder [46]. Different from the MVS branch, only the reference image is fed into the MDE branch. The DPT decoder produces monocular features in a multi-stage manner, which naturally corresponds to the cascade multi-stage in the MVS branch. At stage s, the DPT monocular feature $\mathbf{M}^s$ is fused with the regularized cost volume $\mathbf{R}^s$ through BDF. The BDF-refined monocular feature is decoded into a relative monocular depth map. Meanwhile, the BDF-refined monocular feature is propagated to the next stage in the DPT decoder, allowing multi-view geometric evidence injected at coarser stages to guide subsequent high-resolution monocular feature representations.

Through this design, BridgeMVS enables bidirectional interaction between multi-view and monocular representations at every cascade stage. The MVS branch benefits from monocular structural priors for robust cost volume refinement, while the MDE branch benefits from multi-view geometric constraints for improved relative consistency. The fused representations are progressively propagated from coarse to fine stages, producing both metric MVS depth and relative monocular depth.

3.2 Bidirectional Dynamic Fusion

The proposed Bidirectional Dynamic Fusion (BDF) dynamically bridges the regularized cost volume and the DPT monocular feature through mono-to-cost aggregation and cost-to-mono depth probability fusion. As shown in Fig. 3, at stage s, BDF takes the regularized cost volume $\mathbf{R}^s \in \mathbb{R}^{G^s \times D^s \times H^s \times W^s}$ (where G^s, D^s, and $H^s \times W^s$ are the number of groups, number of depth hypotheses and spatial resolution in the MVS branch) and the corresponding DPT monocular feature $\mathbf{M}^s \in \mathbb{R}^{C_m^s \times H_m^s \times W_m^s}$ (where C_m^s and $H_m^s \times W_m^s$ are the channel dimension and spatial resolution in the MDE branch) as inputs. Here, $\mathbf{R}^s$ encodes multi-view geometric evidence, while $\mathbf{M}^s$ provides strong monocular structural priors.

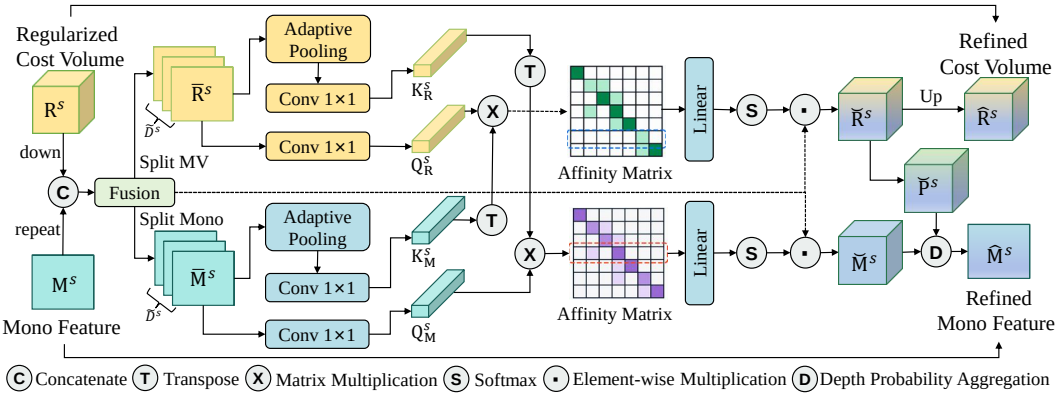


Figure 3: Overview of the proposed bidirectional dynamic fusion module. It first aligns and fuses monocular features and multi-view cost volumes, then performs bidirectional dynamic aggregation, followed by depth-guided feature fusion and residual update to produce refined multi-view cost volumes and monocular features.

Feature Alignment and Fusion. Since the two inputs have different forms and dimensions, we first align them into a shared 2D fusion space. For $\mathbf{R}^s$, it is downsampled by a 3D convolution layer and permuted along the first two dimensions to produce the aligned cost volume $\tilde{\mathbf{R}}^s \in \mathbb{R}^{\tilde{D}^s \times C_v \times \tilde{H}^s \times \tilde{W}^s}$, where $\tilde{D}^s$, C_v and $\tilde{H}^s \times \tilde{W}^s$ are the number of depth hypotheses, the number of channels and the spatial resolution, respectively. For $\mathbf{M}^s$, it is firstly bilinearly resized to $\tilde{H}^s \times \tilde{W}^s$ and projected to C_v channels by a 1×1 2D convolution, and then repeated along the depth dimension to produce the aligned monocular feature $\tilde{\mathbf{M}}^s \in \mathbb{R}^{\tilde{D}^s \times C_v \times \tilde{H}^s \times \tilde{W}^s}$.

In the following fusion and enhancement procedure, *the depth dimension is independently processed* (similar to the batch dimension) since the monocular feature has no depth dimension in fact. The two aligned features are concatenated along the channel dimension and fed into a lightweight fusion module [61]. The fused representation is then split into an MVS representation $\bar{\mathbf{R}}_s$ and an MDE representation $\bar{\mathbf{M}}_s$, which are used for the subsequent bidirectional dynamic aggregation.

Bidirectional Dynamic Aggregation. The dynamic aggregation is inspired by context-mixing dynamic convolution [62], which is implemented with the neighborhood attention [63]. Differently, our bidirectional dynamic aggregation is mutual for both branches. As shown in Fig. 3, the mono-to-cost and cost-to-mono aggregation are equal in their implementation, but in an inverse direction.

Take the mono-to-cost aggregation for example, we first get the query $\mathbf{Q}^s_{\mathbf{R}} \in \mathbb{R}^{\tilde{D}^s \times C_v \times \tilde{H}^s \tilde{W}^s}$ from $\bar{\mathbf{R}}_s$ with a 1×1 2D convolution and a reshape operation. The key $\mathbf{K}^s_{\mathbf{M}} \in \mathbb{R}^{\tilde{D}^s \times C_v \times \tilde{H}^s \tilde{W}^s / P^2}$ is obtained from $\bar{\mathbf{M}}_s$ with an adaptive pooling, 1×1 convolution and a reshape operation, where P is the pooling kernel size. Then, an affinity matrix $\mathbf{A}_{\mathbf{M}\to\mathbf{R}} \in \mathbb{R}^{\tilde{D}^s \times 1 \times \tilde{H}^s \tilde{W}^s \times \tilde{H}^s \tilde{W}^s / P^2}$ is obtained the inner product between $\mathbf{Q}^s_{\mathbf{R}}$ and $\mathbf{K}^s_{\mathbf{M}}$ along the channel dimension. by $\mathbf{A}_{\mathbf{M}\to\mathbf{R}} = (\mathbf{Q}^s_{\mathbf{R}})^\top \mathbf{K}^s_{\mathbf{M}}$. The neighborhood attention matrix $\tilde{\mathbf{A}}_{\mathbf{M}\to\mathbf{R}} \in \mathbb{R}^{\tilde{D}^s \times 1 \times \tilde{H}^s \tilde{W}^s \times K^2}$ is finally produced by feeding $\mathbf{A}_{\mathbf{M}\to\mathbf{R}}$ to a linear layer (followed by a softmax function along the last dimension), where K is the window size for neighborhood attention. The aggregated cost volume $\check{\mathbf{R}}^s \in \mathbb{R}^{\tilde{D}^s \times C_v \times \tilde{H}^s \times \tilde{W}^s}$ is obtained by $\check{\mathbf{R}}^s = \mathrm{NA}(\bar{\mathbf{R}}^s, \tilde{\mathbf{A}}_{\mathbf{M}\to\mathbf{R}})$, where $\mathrm{NA}(\cdot,\cdot)$ is neighborhood attention proposed in [63].

Similarly, the aggregated monocular feature $\check{\mathbf{M}}^s \in \mathbb{R}^{\tilde{D}^s \times C_v \times \tilde{H}^s \times \tilde{W}^s}$ in the cost-to-mono direction can be obtained in a similar way. But differently, the query and key for affinity matrix computation are obtained from $\bar{\mathbf{M}}_s$ and $\bar{\mathbf{R}}_s$, respectively.

Finally, both $\check{\mathbf{R}}^s$ and $\check{\mathbf{M}}^s$ are permuted to the shape of $C_v \times \tilde{D}^s \times \tilde{H}^s \times \tilde{W}^s$, which are denoted as $\breve{\mathbf{R}}^s$ and $\breve{\mathbf{M}}^s$, respectively. In the mono-to-cost direction, monocular structural priors dynamically guide the aggregation of the MVS representation. In the cost-to-mono direction, multi-view geometric evidence dynamically updates the MDE representation. The feature representations are mutually enhanced by each other.

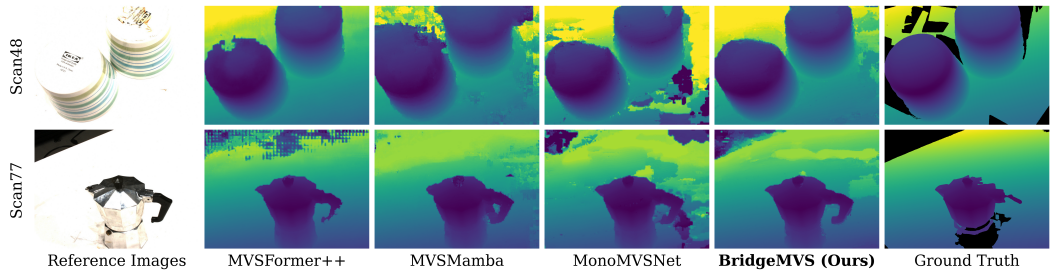


Figure 4: Qualitative comparison of depth maps in most challenging scan48 and scan 77 on DTU. Our method predicted more accurate depth maps in weakly textured and reflection areas.

Depth Probability Fusion and Residual Update. Since the MDE branch does not require the depth dimension, the depth dimension in $\breve{\mathbf{M}}^s$ needs to be reduced. To do this, a depth probability volume $\breve{\mathbf{P}}^s \in \mathbb{R}^{1\times\tilde{D}^s\times\tilde{H}^s\times\tilde{W}^s}$ is obtained by mapping $\breve{\mathbf{R}}^s$ with a lightweight probability projection layer (followed by a softmax activation along the depth dimension). The reduced aggregated monocular feature $\breve{\mathbf{M}}_r^s \in \mathbb{R}^{C_v\times\tilde{H}^s\times\tilde{W}^s}$ is obtained as the element-wise inner product between $\breve{\mathbf{P}}^s$ and $\breve{\mathbf{M}}^s$ along the depth dimension.

Both the $\breve{\mathbf{R}}^s$ and $\breve{\mathbf{M}}_r^s$ are mapped to the original cost volume space and monocular feature space with convolution layers, and used as the residual update, which are added to $\mathbf{R}^s$ and $\mathbf{M}^s$ respectively, producing the enhanced cost volume $\hat{\mathbf{R}}^s \in \mathbb{R}^{G^s\times D^s\times H^s\times W^s}$ and monocular feature $\hat{\mathbf{M}}^s \in \mathbb{R}^{C_m^s\times H_m^s\times W_m^s}$. The enhanced cost volume $\hat{\mathbf{R}}^s$ and monocular feature $\hat{\mathbf{M}}^s$ are used for the depth prediction in stage s, but also propagated to the following stage in their corresponding branches.

3.3 Training Objective

BridgeMVS is supervised at each cascade stage by both the MVS branch and the MDE branch. The MVS branch follows common cascaded multi-stage MVS methods and adopts the cross-entropy loss over the predicted depth probability volume, which provides metric-scale supervision for MVS depth estimation. The MDE branch applies the scale-invariant loss [45] to the relative monocular depth prediction, which constrains monocular depth in a scale-invariant manner and preserves its structural prior. Specifically, let $\hat{\mathbf{P}}^s, \mathbf{P}_{gt}^s \in \mathbb{R}^{D^s\times H^s\times W^s}$ be the predicted probability volume and ground-truth one-hot probability volume for stage s in the MVS branch, and $\hat{\mathbf{D}}^s, \mathbf{D}_{gt}^s \in \mathbb{R}^{H_m^s\times W_m^s}$ be the predicted depth map and ground-truth depth map for stage s in the MDE branch, the overall training objective is defined as:

$$\mathcal{L} = \sum_s \left[\lambda_{mvs}^s \mathcal{L}_{ce}\left(\hat{\mathbf{P}}^s, \mathbf{P}_{gt}^s\right) + \lambda_{mde}^s \mathcal{L}_{si}\left(\hat{\mathbf{D}}^s, \mathbf{D}_{gt}^s\right)\right], \tag{1}$$

where λ_{mvs}^s and λ_{mde}^s are stage-wise loss weights. The metric-scale MVS supervision and relative-scale MDE supervision jointly constrain the coarse-to-fine propagation of fused representations.

4 Experiments

4.1 Datasets and Benchmarks

DTU [36] is an indoor MVS dataset captured under controlled laboratory conditions. BlendedMVS [64] is a large-scale synthetic dataset containing both indoor and outdoor scenes. Tanks-and-Temples (T&T) [37] is a real-world MVS benchmark with complex geometry and diverse acquisition conditions. It includes 8 scenes in the intermediate set and 6 scenes in the advanced set. ETH3D [38] benchmark contains high-resolution indoor and outdoor scenes with large viewpoint variations, and is divided into training and test sets. RobustMVD [65] is a multi-view depth estimation benchmark comprising five datasets: KITTI [66], ScanNet [67], ETH3D [38], DTU [36], and T&T [37].

Table 1: Quantitative results of reconstructed point clouds on the DTU [36] evaluation set with distance metric [mm]. The best , second-best , and third-best results are highlighted with colors, respectively. Methods marked with * use high-resolution images for training.

Methods	**Overall** ↓	Accuracy ↓	Completeness ↓
CasMVSNet [26]	0.355	0.325	0.385
UniMVSNet [71]	0.315	0.352	0.278
MVSTER* [55]	0.313	0.350	0.276
TransMVSNet [16]	0.305	0.321	0.289
DMVSNet [72]	0.305	0.338	0.272
GeoMVSNet [73]	0.295	0.331	0.259
ET-MVSNet [17]	0.291	0.329	0.253
MVSFormer* [33]	0.289	0.327	0.251
GoMVS [34]	0.287	0.347	0.227
RRT-MVS [32]	0.285	0.309	0.261
MVSFormer++* [18]	0.281	0.309	0.252
MVSMamba* [20]	0.280	0.308	0.252
MonoMVSNet [35]	0.278	0.313	0.243
MVSAnywhere [53]	0.735	0.845	0.625
BridgeMVS (Ours)	0.268	0.297	0.239
BridgeMVS* (Ours)	0.262	0.303	0.221

Table 2: Quantitative point cloud results on the T&T [37] benchmark with F-score (%). The best , second-best , and third-best results are marked with colors.

Methods	Intermediate set ↑									Advanced set ↑						
	Mean	Fam.	Fra.	Hor.	L.H.	M60	Pan.	P.G.	Tra.	**Mean**	Aud.	Bal.	Cou.	Mus.	Pal.	Tem.
CasMVSNet [26]	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 [16]	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
MVSTER [55]	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
UniMVSNet [71]	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
ET-MVSNet [17]	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 [33]	66.38	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
DMVSNet [72]	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
GeoMVSNet [73]	65.89	81.64	67.53	55.78	68.02	65.49	67.19	63.27	58.22	41.52	30.23	46.54	39.98	53.05	35.98	43.34
MVSFormer++ [18]	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
GoMVS [34]	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
RRT-MVS [32]	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
MVSMamba [20]	67.67	82.47	72.90	58.55	69.63	65.34	66.88	65.60	59.98	43.32	30.95	49.61	41.04	54.92	36.72	46.67
MonoMVSNet [35]	68.63	82.38	72.89	62.80	70.49	65.79	68.54	65.54	60.59	43.58	30.33	46.76	42.90	56.31	37.28	47.88
BridgeMVS (Ours)	69.47	83.25	73.81	66.80	70.32	66.30	68.88	64.57	61.82	44.80	34.25	49.06	42.86	55.35	36.83	50.44

4.2 Implementation Details

BridgeMVS is implemented in PyTorch [68] and optimized using Adam [69]. Following common practice, we first train and evaluate BridgeMVS on DTU [36]. For DTU training, we use 5 input views at a resolution of 512×640 for 15 epochs, with an initial learning rate of 1e-3. To improve generalization, we further fine-tune the model on a mixture of DTU and BlendedMVS [64], using 11 input views at a resolution of 512×640 for an additional 15 epochs, with an initial learning rate of 4e-4. In addition, following [55, 33, 18, 20], we conduct high-resolution training on DTU using 5 input views at a resolution of 1024×1280 for 10 epochs, with an initial learning rate of 1e-3 (the model is denoted as BridgeMVS*). We use the ViT-B version of Depth Anything V2 [39] as the MDE branch, freezing the encoder and fine-tuning the DPT decoder with a learning rate 0.05× that of the rest of the network. BridgeMVS uses four coarse-to-fine stages with 32-16-8-4 depth hypotheses and 4-4-4-4 grouped feature correlations. We using the dynamic fusion strategy [70] to obtain the final reconstructed point clouds.

4.3 Benchmark Performance

Evaluation on DTU. We evaluate BridgeMVS and BridgeMVS* on DTU [36] evaluation set, using the standard 5 input views at resolutions of 832×1152 and 1152×1600, respectively. We use the official evaluation script to report accuracy, completeness, and their average overall error. As shown in Tab. 1, both BridgeMVS and BridgeMVS* outperform previous state-of-the-art methods by a large

Table 3: Quantitative point cloud results on the ETH3D [38] high-res benchmark with F_1-score [%] at the evaluation threshold of 2 cm. The best, second-best, and third-best results are marked with colors, respectively. Methods marked with † are reproduced by using their open-source code.

Methods	Training Set			Test Set		
	F_1-score ↑	Precision ↑	Recall ↑	F_1-score ↑	Precision ↑	Recall ↑
MonoMVSNet† [35]	70.50	72.96	68.80	76.47	76.66	76.99
MVSTER [55]	72.06	68.08	76.92	79.01	77.09	82.47
MVSMamba† [20]	76.25	77.85	75.14	80.06	80.30	81.55
RRT-MVS† [32]	75.57	76.82	75.00	82.60	82.97	82.61
MVSFormer [33]	73.44	73.62	74.64	82.85	82.23	83.75
MVSFormer++ [18]	-	-	-	82.99	81.88	83.88
EPP-MVSNet [74]	74.00	82.76	67.58	83.40	85.47	81.79
VisMVSNet [75]	72.77	83.32	65.53	83.46	86.86	80.92
Effi-MVS+ [76]	75.28	82.33	70.42	83.62	83.90	83.81
CasDiffMVS [25]	76.75	79.93	75.20	85.11	85.21	85.37
GoMVS [34]	79.16	81.22	77.65	85.91	85.50	86.85
BridgeMVS	79.68	79.44	80.40	86.78	84.80	89.08

Table 4: Quantitative depth estimation results on the RobustMVD [65] benchmark. †: relative depth are alignment with GT using the ROE [40] solver. ‡: trained on eight datasets.

Methods	DTU		T&T		ETH3D		KITTI		ScanNet		Average	
	rel ↓	τ ↑	rel ↓	τ ↑	rel ↓	τ ↑	rel ↓	τ ↑	rel ↓	τ ↑	rel ↓	τ ↑
Depth Anything V2† [39]	5.9	53.2	7.0	50.5	7.9	42.3	7.1	41.2	12.7	24.3	8.1	42.3
BridgeMVS (Mono)†	0.9	95.5	3.0	77.9	4.9	58.9	6.0	45.6	5.5	49.2	4.1	65.4
MonoMVSNet [35]	2.8	81.1	6.5	81.4	7.5	50.0	5.1	58.2	9.7	31.8	6.3	60.5
MVSAnywhere [53]	1.4	90.7	2.7	80.1	**3.6**	**61.5**	4.1	57.8	**4.8**	**53.7**	**3.3**	68.8
BridgeMVS (Ours)	**0.9**	**94.4**	**2.5**	**89.8**	4.9	57.3	**3.7**	**66.5**	6.8	40.3	3.8	**69.7**
MVSAnywhere (full)‡	1.3	95.0	2.1	90.5	3.2	68.0	3.2	68.8	3.7	62.9	2.7	77.0

margin, achieving overall errors of $\mathbf{0.268}$ and $\mathbf{0.262}$, respectively. Qualitative comparisons are shown in Fig. 4, with more results provided in Fig. 5 of Appendix B. Compared with MonoMVSNet [35], which also exploits monocular priors, BridgeMVS improves all metrics and predicts more accurate depth maps, while maintaining nearly the same GPU memory consumption and running time, as shown in Tab. 7. These results demonstrate that bidirectional interaction between monocular features and multi-view cost volumes is more effective than simple unidirectional monocular guidance.

Evaluation on T&T. To assess the generalization capability, we evaluate BridgeMVS on T&T [37] benchmark with 21 input views. Tab. 2 presents the quantitative results. BridgeMVS surpasses all other state-of-the-art methods with mean F-score of **69.47** and **44.80** on both Intermediate and Advanced sets, respectively. Additional qualitative results are shown in Fig. 6 of Appendix B.

Evaluation on ETH3D. To further assess the generalization ability, we evaluate BridgeMVS on the challenging ETH3D [38] high-resolution benchmark using 11 input views at a resolution of 1600×2432. As shown in Tab. 3, BridgeMVS achieves the best F_1-score on both the training and test sets. ETH3D contains high-resolution scenes with large baselines and strong viewpoint variations, posing significant challenges for learning-based MVS methods. Since recent state-of-the-art methods on T&T, such as RRT-MVS [32], MVSMamba [20] and MonoMVSNet [35], have not reported results on ETH3D, we evaluate them using their open-source code and pretrained weights. The clear performance gains demonstrate the strong generalization ability of BridgeMVS.

Evaluation on RobustMVD. Tab. 4 evaluates depth estimation on RobustMVD [65]. We report absolute relative error (rel) and inlier percentage τ with a threshold of 1.03. BridgeMVS (Mono) improves over the baseline Depth Anything V2 [39] across all five datasets, showing that the

Table 5: Structure analysis of Bidirectional Dynamic Fusion (BDF). Gray numbers indicate results from the pre-trained monocular foundation model [39].

Methods (ViT-B)	MVS Branch					MDE Branch	
	Overall ↓	Acc. ↓	Comp. ↓	MAE ↓	RMSE ↓	rel ↓	τ ↑
(a) Base (w/o BDF)	0.292	0.315	0.269	10.66	30.59	8.30	59.6
(b) Mono (frozen) → Cost	0.279	0.312	0.246	7.61	23.11	2.77	71.4
(c) Mono (trainable) → Cost	0.276	0.309	0.243	6.83	21.76	2.15	80.0
(d) Cost → Mono (trainable)	0.289	0.313	0.265	9.04	26.27	0.95	**94.7**
(e) BDF (w/o pretrain)	0.281	0.311	0.251	8.05	23.71	1.16	92.5
(f) BDF (w/o mono loss)	0.273	0.307	**0.239**	7.36	22.56	8.09	30.7
(g) BDF (Ours)	**0.268**	**0.297**	**0.239**	**6.56**	**20.71**	**0.92**	93.6

Table 6: Compatibility study of Bidirectional Dynamic Fusion (BDF) as a plug-in.

Methods (ViT-S)	MVS Branch					MDE Branch		Mem. (MB) ↓	Time (s) ↓
	Overall ↓	Acc. ↓	Comp. ↓	MAE ↓	RMSE ↓	rel ↓	τ ↑		
MVSMamba	0.287	0.314	0.260	9.59	27.78	8.30	59.6	**2090**	**0.11**
+ BDF	**0.275**	**0.291**	**0.259**	**7.43**	**22.84**	**0.93**	**93.6**	2706	0.27
MonoMVSNet	0.278	0.313	0.243	12.94	34.65	8.30	59.6	**2822**	**0.24**
+ BDF	**0.270**	**0.304**	**0.236**	**9.43**	**24.91**	**1.19**	**91.2**	3082	0.31

monocular branch benefits from multi-view geometry transferred through BDF. Under matched DTU+BlendedMVS training data, BridgeMVS outperforms MonoMVSNet [35] and surpasses MVSAnywhere [53] on DTU, T&T, and KITTI, with a higher average τ. MVSAnywhere performs better on ScanNet and ETH3D, where limited camera motion or large viewpoint changes may weaken matching evidence and favor its DPT decoder's use of monocular priors. Although MVSAnywhere (full) achieves better average depth metrics with training on eight datasets, BridgeMVS produces better DTU point clouds (Tab. 1), demonstrating its strength in multi-view reconstruction. More qualitative results appear in Fig. 7 of Appendix B.

4.4 Ablation Study

We conduct ablation studies on the DTU [36] evaluation set using both point cloud and depth metrics. For MVS depth estimation, we report MAE and RMSE. For monocular depth estimation, we report rel and τ. Unless otherwise specified, all ablation experiments use the model trained on DTU at low resolution and are evaluated with 5 input views at a resolution of 832×1152, with all other hyperparameters kept consistent.

Structure Analysis of BDF. Tab. 5 analyzes the structure of the proposed BDF module. Compared with the baseline (a), the one-way Mono→Cost variants (b) and (c) improve MVS reconstruction, showing that monocular structural priors benefit cost volume refinement. Adapting the monocular decoder in (c) further improves the results over the frozen variant in (b). Conversely, Cost→Mono in (d) substantially improves monocular depth estimation but provides only a small MVS gain, suggesting that multi-view evidence effectively refines monocular representations. Combining both directions, the full BDF in (g) improves Overall from 0.276 to 0.268 and monocular rel from 2.15 to 0.92 over the strongest one-way variant (c), yielding the best joint result. Without monocular supervision (f), MVS remains competitive, but monocular rel deteriorates to 8.09. Removing monocular pretraining (e) degrades both branches. These results suggest that pretraining supplies the structural prior, while stage-wise MDE supervision helps maintain informative monocular representations during bidirectional fusion.

Plug-and-play Compatibility. Tab. 6 evaluates the compatibility of BDF as a plug-in module for different cascaded MVS frameworks. Integrating BDF into MVSMamba [20] and MonoMVSNet [35] consistently improves both MVS and MDE metrics, demonstrating that BDF is not limited to a specific baseline. Notably, the runtime increase in MVSMamba mainly stems from introducing the

Table 7: Comparison with MonoMVSNet [35] using different ViT variants.

Methods	Overall↓	Acc.↓	Comp. ↓	MAE ↓	RMSE ↓	Mem. (MB) ↓	Time (s) ↓
MonoMVSNet-S	0.278	0.313	**0.243**	12.94	34.65	2822	**0.24**
BridgeMVS-S	**0.274**	**0.299**	0.249	**7.21**	**21.90**	**2734**	**0.24**
MonoMVSNet-B	0.278	**0.294**	0.262	10.91	30.52	3354	0.35
BridgeMVS-B	**0.268**	0.297	**0.239**	**6.56**	**20.71**	**3132**	**0.34**
MonoMVSNet-L	0.281	0.311	0.251	10.37	29.61	4490	0.65
BridgeMVS-L	**0.266**	**0.295**	**0.237**	**6.32**	**20.01**	**4016**	**0.64**

additional MDE branch, whereas BDF adds only 0.07 s and 260 MB to MonoMVSNet, which already incorporates MDE. Since BDF operates on the regularized cost volume and DPT feature, it can be conveniently inserted into existing cascaded MVS frameworks to enhance both MVS and MDE.

Comparison under Different ViT Variants. Tab. 7 compares BridgeMVS with MonoMVSNet under different monocular ViT variants. BridgeMVS consistently outperforms MonoMVSNet with ViT-S, ViT-B, and ViT-L, showing that BDF is effective across different monocular foundation backbones. Notably, these improvements are achieved without additional runtime overhead. Across different ViT variants, BridgeMVS maintains comparable or lower memory consumption and inference time than MonoMVSNet, demonstrating that BDF is an effective and efficient way to exploit monocular foundation models for MVS reconstruction.

5 Conclusion

In this paper, we propose BridgeMVS, a cascaded multi-stage MVS framework that bidirectionally fuses monocular depth features and multi-view cost volumes. The proposed BDF module serves as a plug-and-play component for existing cascaded MVS frameworks, where monocular structural priors refine MVS cost volumes and multi-view geometric evidence improves monocular representations. Extensive experiments show that BridgeMVS achieves state-of-the-art performance on multiple MVS benchmarks and also improves monocular depth estimation.

Acknowledgments

This work was supported by the Beijing Natural Science Foundation (No. 4262058, No. L257003) and Fundamental Research Funds for the Central Universities (No. FRF-TP-25-033).

References

[1] H. Liu, B. Zou, Q. Liu, H. Yu, Q. Mei, J. Jiang, C. Liu, C. Bi, Z. Wang, X. Zhang, Y. Zhan, J. Chen, and H. Ma, "Unposed-to-3d: Learning simulation-ready vehicles from real-world images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026, pp. 24 843–24 853.

[2] J. Lyu, H. Yu, H. Liu, Q. Mei, H. Ma, and T. Hu, "Spar-gs: Single-view pose-free automobile reconstruction with 3d gaussian splatting," in *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2026, pp. 10 307–10 311.

[3] H. Yu, Q. Liu, H. Liu, J. Jiang, J. Lyu, J. Chen, and H. Ma, "Xyzcylinder: Towards compatible feed-forward 3d gaussian splatting for driving scenes via unified cylinder lifting method," *arXiv preprint arXiv:2510.07856*, 2025.

[4] Z. Yuan, C. Qian, J. Tang, R. Chen, Z. Song, L. Sun, X. Chu, Y. Cai, D. Zhang, and S. Li, "AutoDrive-R^2: Incentivizing reasoning and self-reflection capacity for vla model in autonomous driving," in *International Conference on Learning Representations*, 2026, pp. 97 934–97 955.

[5] Z. Yuan, X. Qu, C. Qian, R. Chen, J. Tang, L. Sun, X. Chu, D. Zhang, Y. Wang, Y. Cai *et al.*, "Video-star: Reinforcing open-vocabulary action recognition with tools," in *International Conference on Learning Representations*, vol. 2026, 2026, pp. 51 445–51 468.

[6] S. Galliani, K. Lasinger, and K. Schindler, "Massively parallel multiview stereopsis by surface normal diffusion," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 873–881.

[7] Q. Xu and W. Tao, "Planar prior assisted patchmatch multi-view stereo," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 516–12 523.

[8] Q. Xu, W. Kong, W. Tao, and M. Pollefeys, "Multi-scale geometric consistency guided and planar prior assisted multi-view stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4945–4963, 2022.

[9] Y. Wang, Z. Zeng, T. Guan, W. Yang, Z. Chen, W. Liu, L. Xu, and Y. Luo, "Adaptive patch deformation for textureless-resilient multi-view stereo," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1621–1630.

[10] C. Ren, Q. Xu, S. Zhang, and J. Yang, "Hierarchical prior mining for non-local multi-view stereo," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3611–3620.

[11] Z. Yuan, D. Zhang, Z. Li, C. Qian, J. Chen, Y. Chen, K. Chen, T. Mao, Z. Li, H. Jiang *et al.*, "Dvp-mvs++: Synergize depth-normal-edge and harmonized visibility prior for multi-view stereo," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[12] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "Mvsnet: Depth inference for unstructured multi-view stereo," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 767–783.

[13] Y. Yao, Z. Luo, S. Li, T. Shen, T. Fang, and L. Quan, "Recurrent mvsnet for high-resolution multi-view stereo depth inference," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5525–5534.

[14] R. T. Collins, "A space-sweep approach to true multi-image matching," in *Proceedings CVPR IEEE computer society conference on computer vision and pattern recognition*. Ieee, 1996, pp. 358–363.

[15] K. T. Giang, S. Song, and S. Jo, "Curvature-guided dynamic scale networks for multi-view stereo," in *International Conference on Learning Representations*, 2022.

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

[17] T. Liu, X. Ye, W. Zhao, Z. Pan, M. Shi, and Z. Cao, "When epipolar constraint meets non-local operators in multi-view stereo," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 18 088–18 097.

[18] C. Cao, X. Ren, and Y. Fu, "Mvsformer++: Revealing the devil in transformer's details for multi-view stereo," in *International Conference on Learning Representations*, vol. 2024, 2024, pp. 39 816–39 837.

[19] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[20] J. Jiang, Q. Liu, H. Liu, H. Yu, L. Wang, J. Chen, and H. Ma, "Mvsmamba: Multi-view stereo with state space model," *Advances in Neural Information Processing Systems*, vol. 38, pp. 144 832–144 859, 2025.

[21] F. Wang, S. Galliani, C. Vogel, and M. Pollefeys, "Itermvs: Iterative probability estimation for efficient multi-view stereo," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 8606–8615.

[22] S. Wang, B. Li, and Y. Dai, "Efficient multi-view stereo by iterative dynamic cost volume," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8655–8664.

[23] J. Jiang, M. Cao, J. Yi, and C. Li, "Di-mvs: Learning efficient multi-view stereo with depth-aware iterations," *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3180–3184, 2024.

[24] J. Jiang and H. Ma, "Efficient multi-view stereo with depth-aware iterations and hybrid loss strategy," *Pattern Recognition*, p. 112500, 2025.

[25] F. Wang, Q. Xu, Y.-S. Ong, and M. Pollefeys, "Lightweight and accurate multi-view stereo with confidence-aware diffusion model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[26] X. Gu, Z. Fan, S. Zhu, Z. Dai, F. Tan, and P. 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*, 2020, pp. 2495–2504.

[27] J. Yang, W. Mao, J. M. Alvarez, and M. Liu, "Cost volume pyramid based depth inference for multi-view stereo," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4877–4886.

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

[29] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *CoRR*, vol. abs/1409.0473, 2014.

[30] J. Liao, Y. Ding, Y. Shavit, D. Huang, S. Ren, J. Guo, W. Feng, and K. Zhang, "Wt-mvsnet: window-based transformers for multi-view stereo," *Advances in Neural Information Processing Systems*, vol. 35, pp. 8564–8576, 2022.

[31] W. Chen, H. Xu, Z. Zhou, Y. Liu, B. Sun, W. Kang, and X. Xie, "Costformer: cost transformer for cost aggregation in multi-view stereo," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 2023, pp. 599–608.

[32] J. Jiang, L. Wang, H. Yu, T. Hu, J. Chen, and H. Ma, "Rrt-mvs: Recurrent regularization transformer for multi-view stereo," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 4, 2025, pp. 3994–4002.

[33] C. Cao, X. Ren, and Y. Fu, "Mvsformer: Multi-view stereo by learning robust image features and temperature-based depth," *Transactions on Machine Learning Research*, 2022.

[34] J. Wu, R. Li, H. Xu, W. Zhao, Y. Zhu, J. Sun, and Y. Zhang, "Gomvs: Geometrically consistent cost aggregation for multi-view stereo," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20 207–20 216.

[35] J. Jiang, Q. Liu, H. Yu, H. Liu, L. Wang, J. Chen, and H. Ma, "Monomvsnet: Monocular priors guided multi-view stereo network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2025, pp. 27 806–27 816.

[36] H. Aanæs, R. R. Jensen, G. Vogiatzis, E. Tola, and A. B. Dahl, "Large-scale data for multiple-view stereopsis," *International Journal of Computer Vision*, vol. 120, pp. 153–168, 2016.

[37] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun, "Tanks and temples: Benchmarking large-scale scene reconstruction," *ACM Transactions on Graphics (ToG)*, vol. 36, no. 4, pp. 1–13, 2017.

[38] T. Schops, J. L. Schonberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger, "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*, 2017, pp. 3260–3269.

[39] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," *Advances in Neural Information Processing Systems*, vol. 37, pp. 21 875–21 911, 2024.

[40] R. Wang, S. Xu, C. Dai, J. Xiang, Y. Deng, X. Tong, and J. Yang, "Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 5261–5271.

[41] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 10 371–10 381.

[42] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[43] T. Dao, "Flashattention-2: Faster attention with better parallelism and work partitioning," in *International Conference on Learning Representations*, vol. 2024, 2024, pp. 35 549–35 562.

[44] A. Bochkovskiy, A. Delaunoy, H. Germain, M. Santos, Y. Zhou, S. Richter, and V. Koltun, "Depth pro: Sharp monocular metric depth in less than a second," in *International Conference on Learning Representations*, vol. 2025, 2025, pp. 75 602–75 637.

[45] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 3, pp. 1623–1637, 2020.

[46] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12 179–12 188.

[47] R. Wang, S. Xu, Y. Dong, Y. Deng, J. Xiang, Z. Lv, G. Sun, X. Tong, and J. Yang, "Moge-2: Accurate monocular geometry with metric scale and sharp details," *Advances in Neural Information Processing Systems*, vol. 38, pp. 35 928–35 959, 2025.

[48] H. Jiang, Z. Lou, L. Ding, R. Xu, M. Tan, W. Jiang, and R. Huang, "Defom-stereo: Depth foundation model based stereo matching," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 21 857–21 867.

[49] J. Cheng, L. Liu, G. Xu, X. Wang, Z. Zhang, Y. Deng, J. Zang, Y. Chen, Z. Cai, and X. Yang, "Monster: Marry monodepth to stereo unleashes power," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 6273–6282.

[50] T. Guan, J. Guo, C. Wang, and Y.-H. Liu, "Bridgedepth: Bridging monocular and stereo reasoning with latent alignment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 27 681–27 691.

[51] B. Wen, M. Trepte, J. Aribido, J. Kautz, O. Gallo, and S. Birchfield, "Foundationstereo: Zero-shot stereo matching," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2025, pp. 5249–5260.

[52] J. Liu, G. Xu, X. Wang, C. Zhang, and X. Yang, "Generalized geometry encoding volume for real-time stereo matching," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 9, 2026, pp. 7179–7187.

[53] S. Izquierdo, M. Sayed, M. Firman, G. Garcia-Hernando, D. Turmukhambetov, J. Civera, O. Mac Aodha, G. Brostow, and J. Watson, "Mvsanywhere: Zero-shot multi-view stereo," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 11 493–11 504.

[54] S. Wang, J. Sun, B. Hou, Q. Wang, B. Fan, B. Li, B. Lu, and Y. Dai, "Spe-mvs: Spatial position encoding enhanced multi-view stereo with monocular depth priors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026, pp. 14 536–14 545.

[55] X. Wang, Z. Zhu, G. Huang, F. Qin, Y. Ye, Y. He, X. Chi, and X. Wang, "Mvster: Epipolar transformer for efficient multi-view stereo," in *European Conference on Computer Vision*. Springer, 2022, pp. 573–591.

[56] S.-W. Kim, H.-K. Kook, J.-Y. Sun, M.-C. Kang, and S.-J. Ko, "Parallel feature pyramid network for object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 234–250.

[57] X. Guo, K. Yang, W. Yang, X. Wang, and H. Li, "Group-wise correlation stereo network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3273–3282.

[58] Q. Xu, W. Su, Y. Qi, W. Tao, and M. Pollefeys, "Learning inverse depth regression for pixelwise visibility-aware multi-view stereo networks," *International Journal of Computer Vision*, vol. 130, no. 8, pp. 2040–2059, 2022.

[59] F. Wang, S. Galliani, C. Vogel, P. Speciale, and M. Pollefeys, "Patchmatchnet: Learned multi-view patchmatch stereo," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14 194–14 203.

[60] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. HAZIZA, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *Transactions on Machine Learning Research*, 2024.

[61] X. Ding, Y. Zhang, Y. Ge, S. Zhao, L. Song, X. Yue, and Y. Shan, "Unireplknet: A universal perception large-kernel convnet for audio video point cloud time-series and image recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 5513–5524.

[62] M. Lou and Y. Yu, "Overlock: An overview-first-look-closely-next convnet with context-mixing dynamic kernels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 128–138.

[63] A. Hassani, S. Walton, J. Li, S. Li, and H. Shi, "Neighborhood attention transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 6185–6194.

[64] Y. Yao, Z. Luo, S. Li, J. Zhang, Y. Ren, L. Zhou, T. Fang, and L. 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*, 2020, pp. 1790–1799.

[65] P. Schröppel, J. Bechtold, A. Amiranashvili, and T. Brox, "A benchmark and a baseline for robust multi-view depth estimation," in *2022 International Conference on 3D Vision (3DV)*. IEEE, 2022, pp. 637–645.

[66] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361.

[67] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5828–5839.

[68] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[69] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[70] J. Yan, Z. Wei, H. Yi, M. Ding, R. Zhang, Y. Chen, G. Wang, and Y.-W. Tai, "Dense hybrid recurrent multi-view stereo net with dynamic consistency checking," in *European conference on computer vision*. Springer, 2020, pp. 674–689.

[71] R. Peng, R. Wang, Z. Wang, Y. Lai, and R. Wang, "Rethinking depth estimation for multi-view stereo: A unified representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8645–8654.

[72] X. Ye, W. Zhao, T. Liu, Z. Huang, Z. Cao, and X. 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*, 2023, pp. 17 661–17 670.

[73] Z. Zhang, R. Peng, Y. Hu, and R. Wang, "Geomvsnet: Learning multi-view stereo with geometry perception," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 508–21 518.

[74] X. Ma, Y. Gong, Q. Wang, J. Huang, L. Chen, and F. Yu, "Epp-mvsnet: Epipolar-assembling based depth prediction for multi-view stereo," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5732–5740.

[75] J. Zhang, S. Li, Z. Luo, T. Fang, and Y. Yao, "Vis-mvsnet: Visibility-aware multi-view stereo network," *International Journal of Computer Vision*, vol. 131, no. 1, pp. 199–214, 2023.

[76] S. Wang, B. Li, and Y. Dai, "Efficient multi-view stereo by dynamic cost volume and cross-scale propagation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 10, pp. 9414–9427, 2024.

[77] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, and D. Novotný, "Vggt: Visual geometry grounded transformer," *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306, 2025.

[78] Y. Wang, J. Zhou, H. Zhu, W. Chang, Y. Zhou, Z. Li, J. Chen, J. Pang, C. Shen, and T. He, "π^3: Scalable permutation-equivariant visual geometry learning," in *International Conference on Learning Representations*, 2026, pp. 10 481–10 497.

[79] V. Leroy, Y. Cabon, and J. Revaud, "Grounding image matching in 3d with mast3r," in *European conference on computer vision*. Springer, 2024, pp. 71–91.

[80] H. Lin, S. Chen, J. H. Liew, D. Y. Chen, Z. Li, Y. Zhao, S. Peng, H. Guo, X. Zhou, G. Shi, J. Feng, and B. Kang, "Depth anything 3: Recovering the visual space from any views," in *International Conference on Learning Representations*, 2026, pp. 141 261–141 285.

A More Experimental Results

A.1 Comparison with Feed-forward Reconstruction Models

Tab. 8 compares BridgeMVS with recent feed-forward reconstruction models, including VGGT [77], π^3 [78], MASt3R [79], and Depth Anything 3 [80], using the official DTU [36] point cloud metrics. For our runs of VGGT, π^3, and Depth Anything 3, predicted depths are aligned to GT through least-squares fitting and fused using GT camera poses, following the VGGT evaluation protocol. "Known GT Camera" indicates whether GT cameras are used to derive depth or geometry before fusion, rather than whether GT poses are used during fusion. MASt3R obtains geometry by triangulating feature correspondences using GT cameras. BridgeMVS achieves the lowest listed Overall error of 0.268 in its calibrated MVS setting. These results demonstrate its effectiveness for calibrated reconstruction, while differences in camera assumptions and the GT-depth alignment used for feed-forward predictions make this comparison additional quantitative context rather than a strictly equivalent evaluation.

Table 8: Comparison with feed-forward reconstruction models of point clouds on DTU.

Methods	Known GT Camera	Overall ↓	Accuracy ↓	Completeness ↓
VGGT (paper) [77]	×	0.382	0.389	0.374
VGGT (our run) [77]	×	0.525	0.634	0.416
π^3 (our run) [78]	×	0.921	1.025	0.817
Depth Anything 3 (our run) [80]	×	1.235	1.432	1.038
MASt3R (paper) [79]	✓	0.374	0.403	0.344
Depth Anything 3 (our run) [80]	✓	1.210	1.410	1.010
BridgeMVS (Ours)	✓	**0.268**	**0.297**	**0.239**

A.2 Comparison with Alternative Fusion Strategies

We compare BDF with static bidirectional fusion and BridgeDepth-style Iterative Bidirectional Alignment (IBA) in Tab. 9. All variants use the same BridgeMVS backbone, training schedule, and DTU evaluation setting (5 views at 832×1152 resolution).

Table 9: Comparison of alternative fusion strategies on DTU. MDE metrics are reported after GT alignment of the auxiliary output, which requires multi-view inputs. Params counts only the parameters of the fusion modules.

Fusion Strategies	MVS branch					MDE branch		Model Efficiency		
	Overall ↓	Acc. ↓	Comp. ↓	MAE ↓	RMSE ↓	rel ↓	τ ↑	Time ↓	Mem. ↓	Params ↓
Concat, $C \times D$	0.278	0.289	0.268	6.78	21.32	2.03	82.1	**0.27**	3146	1.06
Add, $C \times D$	0.283	**0.288**	0.278	6.81	21.23	2.07	81.4	**0.27**	2974	1.03
Concat, per-depth	0.277	0.293	0.260	7.01	22.00	2.04	82.1	**0.27**	3104	0.87
Add, per-depth	0.279	0.300	0.257	7.15	22.48	2.06	82.1	**0.27**	**2904**	**0.85**
IBA [50]	0.276	0.290	0.262	7.95	23.55	1.21	90.5	0.50	13968	0.90
BDF (Ours)	**0.268**	0.297	**0.239**	**6.56**	**20.71**	**0.92**	**93.6**	0.34	3132	0.94

Static bidirectional fusion. We construct four static baselines by combining two cost-volume organizations with concatenation or element-wise addition. In the $C \times D$-folded variants, the cost-channel and depth-hypothesis dimensions are merged into a single feature dimension. The resulting

cost volume and the monocular feature are projected to the same dimension before fusion. In the per-depth variants, the depth dimension is retained, and the projected monocular feature is fused with each projected cost hypothesis. Each shared fused representation generates updates for both branches without input-dependent affinity weights. BDF achieves best performance than all four variants with comparable memory usage and runtime. The improvements across both branches support the benefit of selectively incorporating cross-branch evidence through direction-specific adaptive aggregation.

BridgeDepth-style IBA. We replace BDF with a BridgeDepth-style IBA module using local bidirectional cross-attention. With similar fusion parameters, BDF improves Overall from 0.276 to 0.268, MVS depth MAE from 7.95 to 6.56, and auxiliary monocular relative error from 1.21 to 0.92. It also reduces runtime from 0.50 s to 0.34 s and memory usage from 13,968 MB to 3,132 MB. This comparison concerns the transplanted fusion mechanisms rather than the complete native pipelines: IBA performs alignment at a single post-DPT stage, whereas its adaptation here operates on full MVS depth volumes at all four cascade stages, substantially increasing memory requirements. BDF instead uses cross-branch affinities to guide aggregation of the target branch's own values, supporting multi-stage interaction without direct cross-branch value transfer over all depth hypotheses.

B More Visualization Results

Fig. 5 presents additional depth map comparisons with MonoMVSNet [35] and Depth Anything V2 [39] on the DTU [36] dataset. Benefiting from the bidirectional dynamic fusion between the MVS and MDE branches, BridgeMVS and BridgeMVS (Mono) mutually enhance each other and produce more accurate multi-view and monocular depth predictions. Fig. 6 further shows that our method achieves better accuracy and completeness in challenging regions on the Tanks-and-Temples benchmark [37]. Fig. 7 shows qualitative results on RobustMVD [65], where BridgeMVS (Mono) benefits from multi-view geometric evidence provided by the MVS branch and produces more accurate monocular depth maps with lower relative errors.

C Limitations

BridgeMVS remains subject to limitations of both monocular prediction and calibrated MVS. First, because the Cost-to-Mono update relies on the regularized cost volume, the enhanced monocular output requires multi-view inputs and is not a standalone single-image depth estimator. Second, BDF relies on complementary MDE and MVS cues for mutual enhancement. When both are unreliable, as can occur with complex materials or depth boundaries, bidirectional updates may propagate errors because BDF does not explicitly estimate cue reliability. Third, BridgeMVS assumes known intrinsics/extrinsics and an initial metric depth range, as required by its calibrated plane-sweep MVS backbone. Relative monocular depth alone cannot provide metric bounds. Future work will investigate uncertainty-aware fusion, robustness to perturbations in the required camera parameters, and using monocular depth priors to anchor the initial MVS depth range.

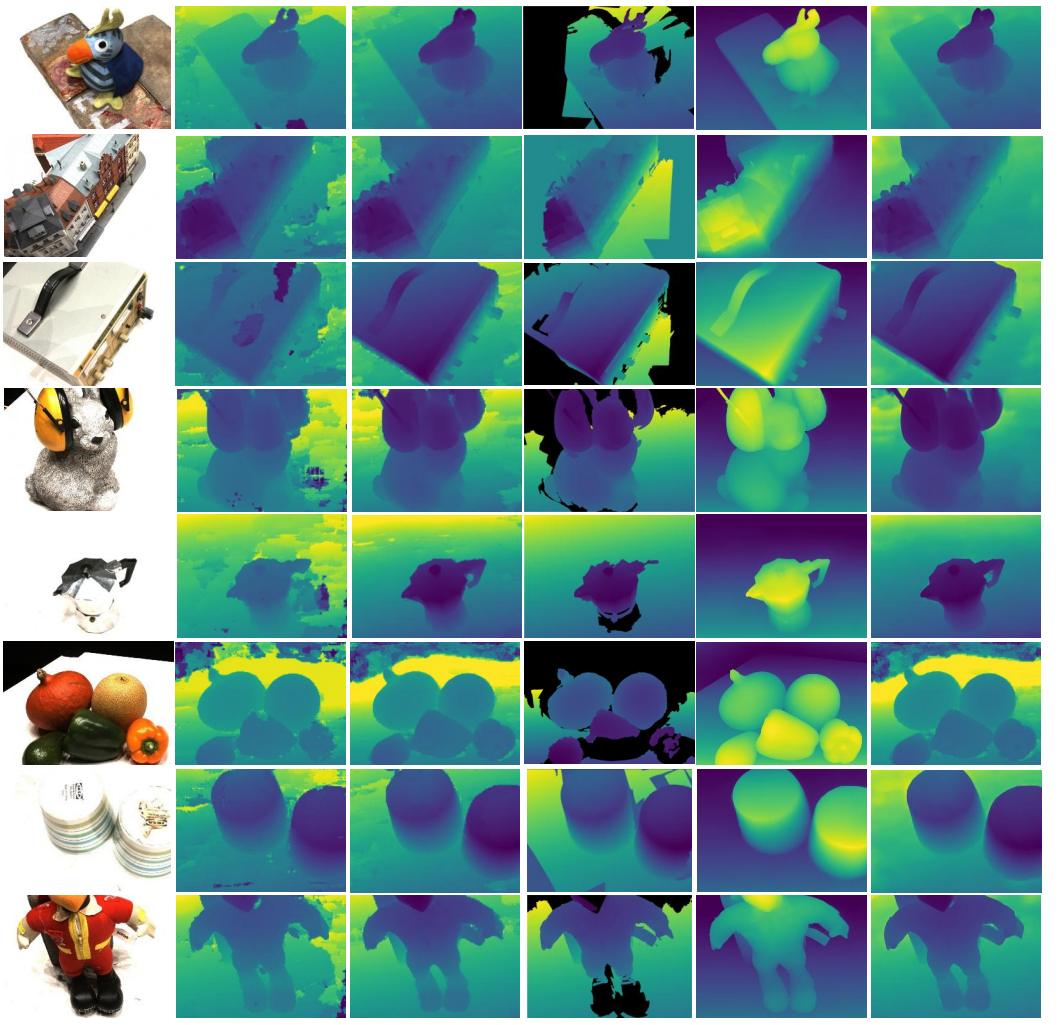


Figure 5: Additional depth maps visualization comparing the MonoMVSNet, BridgeMVS, Ground Truth, Depth Anything V2 (DAV2), and BridgeMVS (Mono) on the DTU [36] dataset.

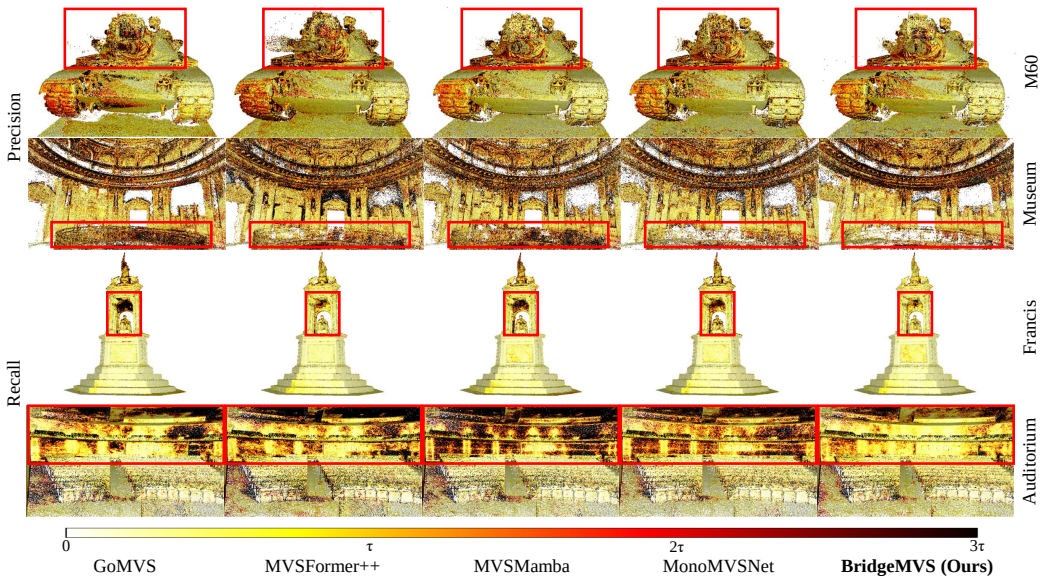


Figure 6: Qualitative comparison of reconstructed point clouds on the Tanks-and-Temples benchmark. Brighter regions indicate lower reconstruction errors under the corresponding distance threshold τ.

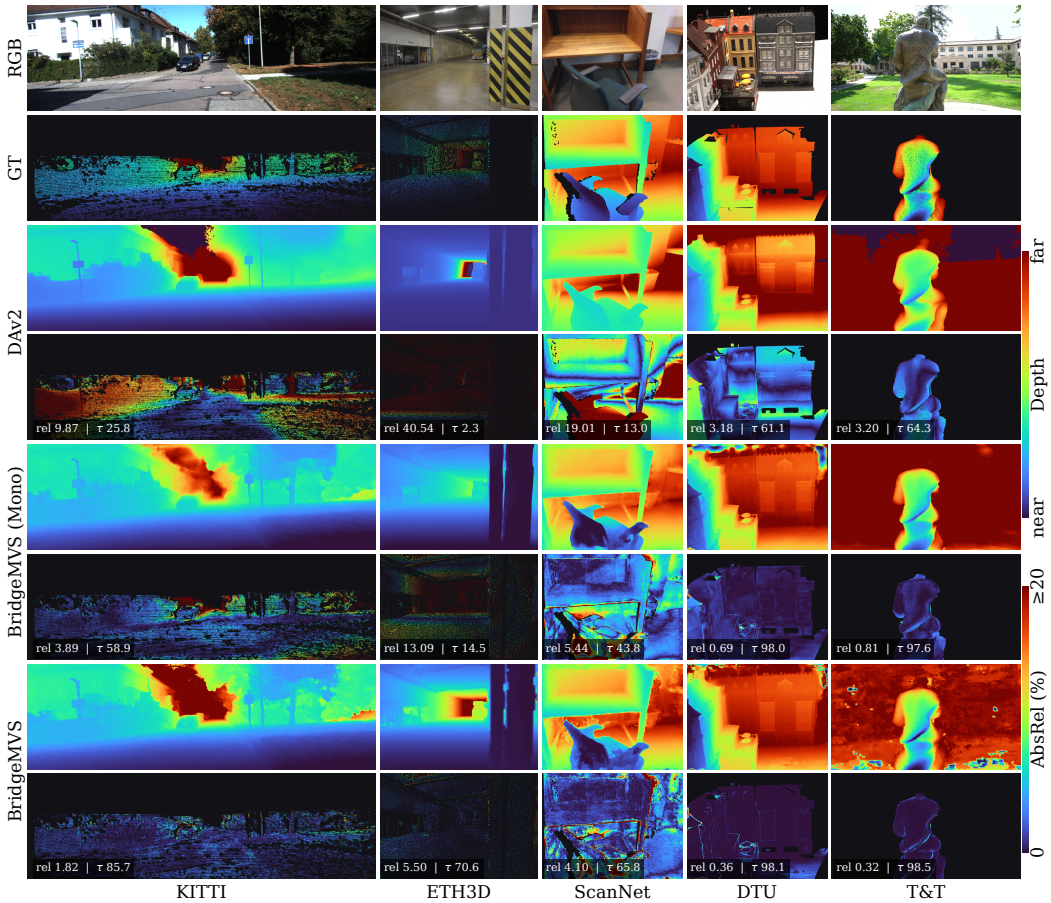


Figure 7: Qualitative comparison of depth maps and absolute relative error maps on the Robust-MVD [65] benchmark. For each method, the first row shows the predicted depth maps and the second row shows the corresponding error maps.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes. The contributions stated in Sec. 1 are directly supported by the theoretical analysis in Sec. 3 and the experiments in Sec. 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes. The limitation concerning the multi-view dependency of our monocular depth estimation branch during the inference stage is explicitly discussed and analyzed in Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: This paper does not include theoretical results; it focuses on the architectural design and empirical evaluation of the BridgeMVS framework.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Yes. All necessary implementation details are fully described in Sec. 4.2, and our code along with pre-trained models will be publicly released upon publication.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The code and pre-trained models are not included in this anonymous submission, but they will be fully open-sourced upon publication.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Yes. All critical experimental configurations—including the multi-stage training curriculum, cascade MVS parameters (e.g., depth hypotheses), and the specific setup for the monocular ViT backbone—are explicitly detailed in Sec. 4.2.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We report metrics based on the average results of multiple experiments.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: We use NVIDIA RTX A6000 GPUs for tranining and NVIDIA RTX 3090 for evalution.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This paper adheres to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: This work will have an impact on the field of MVS.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: There are no security issues in the implementation of this work.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets (e.g., DTU) and code bases are properly cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: This work does not introduce any new datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The work does not involve crowdsourcing or human participants.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The study does not involve human subjects or sensitive data requiring IRB.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs were not used as part of the core methodology in this work.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.