Real-World Multi-View Stereo via Learning RGB-D Structural Consistency from Depth Super-Resolution

Yimei Liu, Jingchao Cao, Hao Fan, Junyu Dong, *Member, IEEE*, Sheng Chen, *Life Fellow, IEEE*

***Abstract*—Learning-based Multi-View Stereo (MVS) methods, typically reliant on cascaded cost volume formulations, perform well on small-scale scenes. However, as the depth range of captured images becomes broader and more varied, the coarse-to-fine d epth s ampling p rocess, w hich d epends s olely o n feature matching, is increasingly prone to local optima. Despite recent advancements in feature representation, depth sampling patterns, and cost aggregation techniques, challenges related to model generalization and computational efficiency persist. In this paper, we propose SR-MVSNet, a novel framework that integrates multi-view feature matching and RGB-D cross-modal structural consistency learning to achieve high-quality 3D reconstruction. Our approach begins with the construction of Low-Resolution (LR) cost volumes for initial LR depth estimation, which are then enhanced to full-resolution via a tailored uncertainty-aware guided depth super-resolution module. To ensure cross-view consistency, the depth maps undergo further refinement through multi-view feature matching. By avoiding high-resolution cost volume processing, our framework improves depth estimation robustness and efficiency. A dditionally, w e i ntroduce an iterative depth fusion post-processing strategy during inference to improve reconstruction in ambiguous matching regions, a critical challenge for MVS methods. Experiments show that our method achieves top-3 performance on the DTU and Tanks & Temples datasets and ranks first o n t he E TH3D dataset. Furthermore, it uses significantly fewer GPU resources than most high performing methods, offering a favorable trade-off between reconstruction quality and computational efficiency.**

***Index Terms*—Multi-view stereo, depth estimation, guided depth super-resolution, 3D reconstruction.**

I. Introduction

Multi-View Stereo (MVS) aims to reconstruct 3D structure of objects or scenes using images captured from multiple known camera viewpoints. Traditional MVS methods [1], [2] have relied on hand-crafted image features and matching metrics, allowing accurate reconstruction in well-textured, ideal Lambertian regions but adversely impacted by occlusions and varying lighting conditions. To address these limitations, recent deep learning-based approaches [3]–[6] employ Convolutional Neural Networks (CNNs) to extract high-level image features that enable more reliable matching, leading to improved performance on various MVS benchmarks [7]–[10].

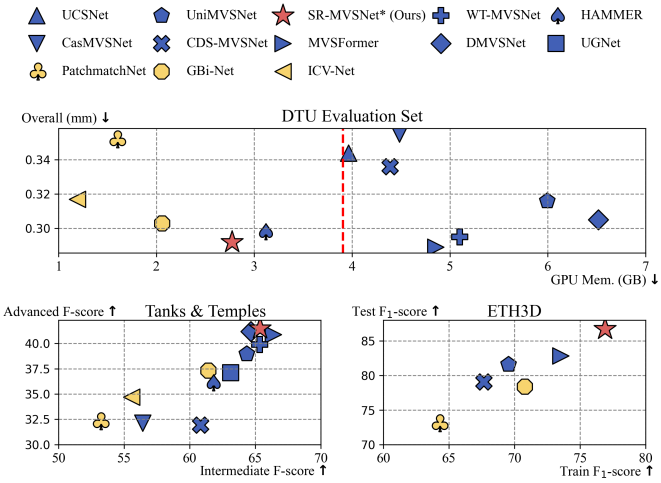


Fig. 1: Comparison with recent lightweight [11]–[13] and coarse-to-fine [14]–[22] methods on the DTU, Tanks & Temples, and ETH3D benchmarks [7]–[9]. Our approach demonstrates competitive reconstruction performance while utilizing less GPU memory than most existing methods.

Learning-based MVS approaches typically follow the cost volume pipeline [3], which constructs 3D cost volumes using generated depth candidates to capture feature matching between reference and source views. These volumes are then processed by 3D CNNs for cost aggregation, enabling depth map inference. A straightforward way to improve depth estimation accuracy is to increase the number of depth candidates (e.g., [23], [24] use 256 candidates). However, this significantly increases memory usage and slows inference. To address this trade-off, previous methods [14], [15] have employed cascaded cost volumes, progressively increasing the feature map resolution while reducing the number of depth candidates, focusing on a narrower range around the previously estimated depths. This coarse-to-fine strategy offers a balance between accuracy and efficiency and has been widely adopted. Further performance improvements and memory optimization have been achieved through advanced feature representations [17], [18] for more robust feature matching and through optimized cost aggregation modules [19] as alternatives to 3D CNNs. Additionally, several studies have focused on refining pixel-wise depth candidate sampling to create more efficient cascaded cost volumes [11]–[13], [20], with some incorporating iterative processes to reduce memory consumption.

Despite these advancements, cascaded cost volume methods [11], [12], [14]–[22], [25] often struggle to balance

This work was supported in part by National Science Foundation of China (Grant No. 42106193, 41927805), Fundamental Research Funds for the Central Universities (No. 202461010). (*Corresponding authors*: Junyu Dong, Jingchao Cao.)

Y. Liu, J. Cao, H. Fan and J. Dong are with the Department of Information Science and Technology, Ocean University of China, Qingdao 266100, China (emails: liuyimei@stu.ouc.edu.cn, caojingchao@ouc.edu.cn, fanhao@ouc.edu.cn, dongjunyu@ouc.edu.cn).

S. Chen is with School of Electronics and Computer Science, University of Southampton, Southampton SO171BJ, UK, and also with the Department of Information Science and Technology, Ocean University of China, Qingdao 266100, China (email: sqc@ecs.soton.ac.uk).

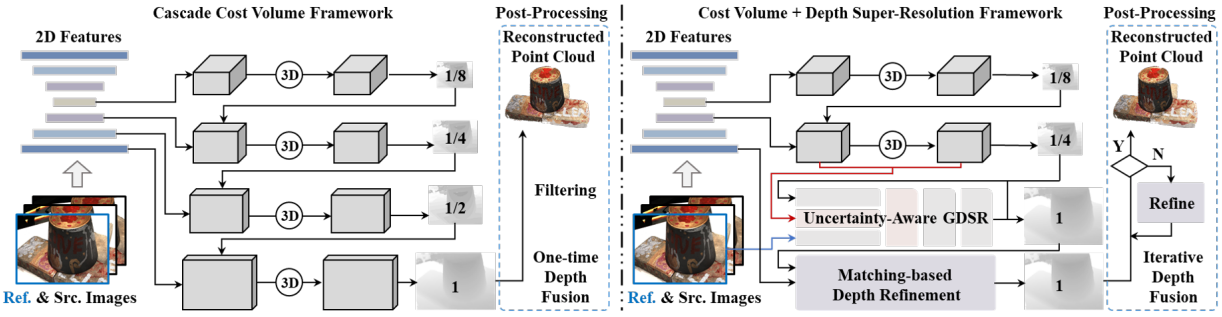


Fig. 2: The left and right diagrams illustrate the key differences between conventional cascaded MVS methods and our proposed framework. The left diagram represents standard cascaded MVS approaches, which build cost volumes across multiple scales for depth estimation and perform a one-time geometric consistency check during post-processing for depth fusion. The right diagram depicts our method, which replaces HR cost volumes with an uncertainty-aware Guided Depth Super-Resolution (GDSR) module and a matching-based depth refinement module. Additionally, we introduce an iterative depth fusion strategy to enhance reconstruction quality, particularly in regions with ambiguous matching issues.

performance with memory usage. Most of these methods require at least 4000 MB of memory to estimate depth at resolutions around 1200×1600, as depicted in Fig. 1. Furthermore, they often face challenges when handling large-scale scenes in benchmarks such as Tanks & Temples (TnT) [8] and ETH3D [9], particularly with fewer images per scene. Previous studies have shown that GPU memory usage and runtime increase cubically with the resolution of the cost volume, rendering full-resolution cost volume processing and depth estimation highly computationally expensive. Besides, as the scene scale expands and the image number decreases, the depth range for each view becomes broader and varied. This makes it challenging for the progressively narrowed depth sampling in the cascaded architecture to consistently cover the ground truth accurately. Meanwhile, ambiguous matching situations become severe. Both negative aspects lead to degradation in reconstruction performance. Then, a motivating question arises: Is there an alternative approach to achieving high-quality, full-resolution depth estimation that minimizes the negative impact of unreliable matching results while maintaining algorithmic efficiency?

Our intuitive idea is to achieve depth estimation by leveraging not only multi-view feature matching results but also the semantic and structural consistency between RGB-D modalities. To this end, we propose a novel "cost volume + depth super-resolution" framework (as shown in Fig. 2). This involves constructing Low-Resolution (LR) cost volumes to estimate the LR depth map, which is subsequently super-resolved to full-resolution using the corresponding full-resolution RGB image as a structural guide. Compared to directly handling High-Resolution (HR) cost volumes, our Guided Depth Super-Resolution (GDSR) manner is less sensitive to variations in depth range and operates more efficiently in 2D space. However, existing GDSR methods [26]–[28] cannot be directly applied to our MVS task, as they typically assume that the input LR depth map is from a single view and is only affected by reduced spatial resolution and noise, with regular sampling and no outliers. More importantly, these methods focus on exploring structure-consistent content between the LR depth map and the HR guidance image without considering multi-view geometric constraints, as this falls outside their research scope. In contrast, LR depths derived from cost volumes in our framework are prone to outliers due to ambiguous matching, and the enhanced full-resolution depth map needs to consider critical multi-view feature matching consistency.

To address the issue of handling LR outliers and ambiguity, we propose an uncertainty-aware GDSR approach for matching-based LR depth super-resolution, extending the classic GDSR method, AHMF [27]. Our approach is driven by two key modules: a Cost Feature Extraction (CFE) module and an Uncertainty-aware Multi-modal Fusion (UMF) module. The CFE captures unreliable LR estimates by applying entropy to the cost volume along the depth dimension, effectively activating cost features in regions of high uncertainty. The UMF module then integrates depth, cost and guidance features into an uncertainty-aware mutual structure feature representation, facilitating RGB-D structurally consistent HR depth prediction. To further enhance this learning process, we incorporate a confidence mask in the loss function. This mask encourages larger residual corrections in uncertain regions where cost features are activated, while preserving reliable estimates in well-matched areas, ensuring a balance between correction and stability.

To ensure geometric consistency, we propose a matching-based refinement module that involves explicit multi-view geometric constraints to further optimize the obtained depth maps. Specifically, refinement is performed through depth self-reintegration of learnable neighboring information, weighted by their feature matching scores.

Furthermore, we propose an iterative depth fusion post-processing strategy (as shown in Fig. 2) to improve the reconstruction performance of ambiguous matching regions. Our post-processing alternates between conventional geometric consistency filtering and filtered depth self-reintegration, a simplified, non-parametric version of our previous matching-based refinement step. By progressively filtering outliers and reintegrating cross-validated points across iterations, our approach generates accurate and complete reconstruction.

Our main contributions are summarized as follows:

- **A "cost volume + depth super-resolution" framework.**

We introduce an uncertainty-aware guided depth super-resolution module to replace the processing of HR cost volumes. By integrating feature matching with RGB-D structural consistency learning, our method enhances matching-based LR depth maps to full resolution, improving both robustness and efficiency of depth estimation.

- **A matching-based depth refinement module.** We propose a feature matching-based depth self-reintegration procedure to ensure multi-view consistent final depth estimation.
- **An iterative depth fusion post-processing strategy.** We alternate between conventional geometric consistency filtering and simplified filtered depth self-reintegration to improve reconstruction performance in regions with ambiguous matching issues.

Experimental results demonstrate that our approach achieves competitive reconstruction performance on the DTU, TnT, and ETH3D benchmarks [7]–[9], while reducing memory usage compared to most learning-based methods [11], [12], [14]–[22], [25]. This highlights the advantages of our approach in terms of both generalization and efficiency (see Fig. 1).

II. Related Work

A. Learning-Based MVS

The advent of deep CNNs has greatly advanced MVS 3D reconstruction. The seminal work MVSNet [3] introduced an end-to-end pipeline combining 3D cost volume formulation with 3D CNN regularization. However, the high computational and memory demands limit input resolution. To address this, previous studies like CasMVSNet [15], UCSNet [14], and NP-CVP-MVS [29] introduced cascaded cost volumes to reduce memory usage. Recent works have focused on enhancing feature representation [17], [18] [30] and optimizing cost aggregation [19], [31] [32]. CDS-MVS [17] enhanced feature representation by computing normal curvature along the epipolar line. MVSTR [30] introduced intra-view and cross-view Transformer modules to improve 3D-consistent feature learning across multiple views. MVSFormer [18] leveraged pre-trained Vision Transformer (ViT) models to facilitate feature extraction in the MVS task. MVSTER [31] and WT-MVSNet [19] designed epipolar-guided, window-based cost transformers to generate more complete and smoother probability volumes. Most recently, GoMVS [32] proposed a geometrically consistent cost aggregation process that effectively integrated adjacent pixel costs, leading to improved depth estimation accuracy.

Some other works have employed pixel-wise depth sampling to formulate more fine-grained cost volumes [11], [12], [20], [25], [33]–[35]. PatchmatchNet [12] introduced depth propagation by sampling from learnable neighboring locations but lacked consideration of the implicit depth distribution within scenes. DS-PMNet [33] addressed this with the DeformSampler, which learns a distribution-sensitive sample space. NR-MVSNet [34] generated depth candidates from neighboring pixels with similar normals. UGNet [20] and ARAI-MVSNet [35] proposed uncertainty-guided sampling to adaptively adjust the depth search range, creating more compact cost volumes. GBi-Net [11] introduced a generalized binary search strategy for efficient depth candidate generation. Building upon this, ICV-Net [13] integrated a dense-to-sparse search mechanism with identity cost volumes, further reducing memory overhead and showing stronger robustness in the early stages.

To reduce GPU memory consumption, several methodss [11], [12], [25] adopted iterative processes that sample a small number of high-probability depth candidates in each iteration to formulate cost volume at the same size multiple times. Although this strategy significantly reduces computational demand, the extremely limited number of depth candidates sampling inevitably sacrifices generalization to large scenarios to some extent.

Beyond improving depth estimation, several works focus on the quality of the reconstructed point cloud. DMVS-Net [21] proposed predicting two depth values per pixel and selecting the final value using a checkerboard-shaped strategy, thus reducing errors from interpolation during the multi-view fusion post-processing step. Other studies refine the widely used one-time photometric and geometric filtering strategy from MVSNet [3], which requires multiple heuristic hyper-parameters. D^2HC-RMVSNet [36] proposed a dynamic consistency filtering strategy for generating more accurate, reliable dense points. More recently, HAMMER [22] reduced manual parameter tuning by learning an entropy-based filtering mask combined with two-view geometric verification.

In contrast to existing cascade-based MVS methods that rely entirely on cost volume formulation for full-resolution depth estimation, our framework (Fig. 2) propose an uncertainty-aware GDRS module designed to incorporate mutual structural cues from RGB-D images for recovering high-quality, full-resolution depth maps. By avoiding the construction of HR cost volumes with narrow sampling depth ranges, our method demonstrates advantages in handling images with diverse depth ranges while maintaining lower resource demands. Additionally, unlike conventional single-pass depth fusion post-processing, our iterative fusion strategy improves the reconstruction of ambiguous matching regions by reintegrating cross-validated points across multiple iterations until convergence.

B. Guided Depth Super-Resolution (GDSR)

GDSR utilizes the HR color image to enhance the spatial resolution of the corresponding LR depth map. It has been primarily applied to depth data from active sensors, i.e., consumer-level time-of-flight cameras or structure light scanners, and aligned color imagery from internally fixed digital cameras. Various approaches have been proposed over the last few years, including filtering-based [37], [38], optimization-based [39]–[41], and dictionary learning [42], [43] and deep learning [27], [28], [44]–[46] techniques (see survey [26] for more details). Different from color image super-resolution methods that adopt a single-branch network, most deep learning-based GDSR methods [27], [28], [44] utilize dual-branch architectures to extract depth and color features independently and then fuse them to rebuild HR depth maps. However, naive multi-modal feature fusion by

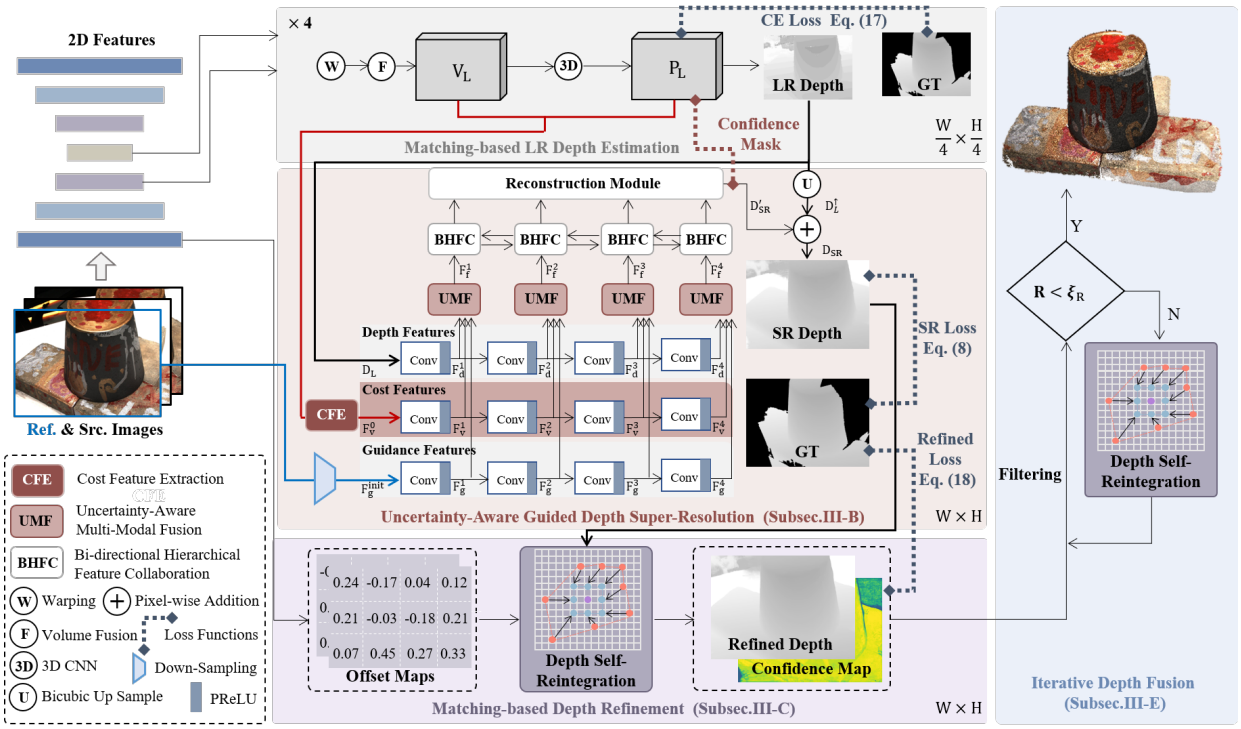


Fig. 3: The overall architecture of SR-MVSNet. LR cost volumes are constructed for matching-based LR depth estimation at $W/4 \times H/4$. The last cost volume, probability volume, LR depth map, and HR reference image are then input into the uncertainty-aware guided depth super-resolution module to predict the full-resolution depth map at $W \times H$. The obtained depth map is subsequently refined through the matching-based refinement module, which leverages multi-view feature consistency. Finally, the multi-view depth maps are fused into the final point clouds using the proposed iterative depth fusion strategy.

concatenation or multiplication is subject to texture-copying artifacts since features from different modalities are equally treated. Recent works incorporated additional components to improve fusion and suppress artifacts, such as residual learning strategy [44], deformable convolutions [45], and co-structural feature exploration [28]. The state-of-the-art AHMF [27] adaptively selects and effectively fuses modalities through a critical multi-modal attention-based fusion module (MMAF) and further explores the complementarity of multi-level fused features to realize HR depth reconstruction. However, existing GDSR methods focus on improving spatial resolution and assume LR inputs are regularly sampled without outliers. In the context of MVS, we extend AHMF [27] to address the uncertainty inherent in matching-based LR depth. The proposed uncertainty-aware guided depth super-resolution module seamlessly integrates with the MVS task, facilitating efficient and reliable full-resolution depth estimation.

III. Methodology

In this section, we introduce the detailed structure of the proposed SR-MVSNet. We begin with an architectural framework overview in Subsection III-A, which also briefly introduces matching-based LR depth estimation from GBi-Net [11]. The following subsections elaborate on three key modules that enhance baseline performance:

- The uncertainty-aware GDSR module (Subsection III-B), which is designed to predict full-resolution depth map from matching-based LR depth map.
- The matching-based depth refinement module (Subsection III-C), which refines the depth map based on explicit multi-view geometric constraints.
- The iterative depth fusion post-processing strategy (Subsection III-E), which is advantageous for complete and accurate 3D reconstruction in ambiguous matching regions.

Subsection III-D defines the loss functions used for model training. These contributions collectively enhance the robustness of HR depth estimation while balancing algorithmic efficiency.

A. Network Overview

The overall architecture of SR-MVSNet is shown in Fig.3. Given a reference image $\mathbf{I}_0$ and several source images $\{\mathbf{I}_i\}_{i=1}^{N}$, with $\mathbf{I}_i \in \mathbb{R}^{3\times W\times H}$, extrinsic transformation from the reference to the source views $\{\mathbf{T}_i\}_{i=1}^{N}$, intrinsic matrixs $\{\mathbf{K}_i\}_{i=0}^{N}$. SR-MVSNet aims to estimate the reference depth map $\mathbf{D}_0 \in \mathbb{R}^{W\times H}$ and fusing multi-view depth maps to reconstruct scene 3D point cloud.

First, we conduct matching-based LR depth estimation by iteratively constructing LR cost volumes four times using generalized binary searched depth candidates [11]. Following standard cost volume construction and regularization procedures, including group-wise correlation, pixel-wise weighted fusion, and 3D-CNN, we derive the LR depth label/map $\mathbf{D}_L$, cost volume $\mathbf{V}_L \in \mathbb{R}^{C\times D\times \frac{W}{4}\times \frac{H}{4}}$, and probability volume $\mathbf{P}_L \in \mathbb{R}^{D\times \frac{W}{4}\times \frac{H}{4}}$ from the final iteration, with $C = 32$ and $D = 4$, denotes the feature channel number and depth candidate number, respectively. Our LR depth estimation phase

consumes approximately half the memory compared to GBi-Net, with detailed parameter settings provided in Appendix A. Subsequently, the proposed uncertainty-aware GDSR and matching-based depth refinement modules are employed to achieve full-resolution depth estimation while ensuring multi-view consistency. Although our method ultimately outputs depth maps at the same resolution as mainstream MVS approaches [3], [11], [18], the integration of RGB-D cross-modal feature interaction and depth self-reintegration modules strengthens both robustness and efficiency in depth estimation. Finally, the estimated multi-view depths undergo iterative fusion to produce dense 3D point clouds.

B. Uncertainty-aware GDSR

The proposed uncertainty-aware GDSR module predicts full-resolution depth from matching-based LR inputs, guided by HR image and uncertainty cues from LR cost volume. As shown in Fig. 3, the input data includes HR reference image $\mathbf{I}_0 \in \mathbb{R}^{3\times W\times H}$, LR cost volume $\mathbf{V}_L \in \mathbb{R}^{C\times D\times \frac{W}{4}\times \frac{H}{4}}$, LR probability volume $\mathbf{P}_L \in \mathbb{R}^{D\times \frac{W}{4}\times \frac{H}{4}}$, and LR depth $\mathbf{D}_L$. We adopt the GDSR model AHMF [27] as our baseline, with upscale factor fixed at 4. Our uncertainty-aware GDSR follows the four-step structure: (1) multi-modal feature extraction, (2) multi-modal feature fusion, (3) bi-directional hierarchical feature collaboration (BHFC), and (4) full-resolution depth reconstruction. In our framework, we enhance the first two steps to handle uncertainty in the input LR depth. The latter two steps, BHFC and final depth reconstruction, remain consistent with AHMF [27]. For brevity, these two steps are not elaborated in the methodology. The key improvements in our uncertainty-aware GDSR module, highlighted in red in Fig. 3, are summarized as follows:

- A novel **Cost Feature Extraction (CFE)** module and an associated cost branch are added in the multi-modal feature extraction step to explicitly model the uncertainty in the LR depth.
- We replace the core multi-modal feature fusion step of AHMF with our **Uncertainty-aware Multi-modal Fusion (UMF)** module, which more effectively integrates RGB and depth features by incorporating uncertainty.
- During training, we normalize the HR RGB image and LR depth map to $[0, 1]$ as inputs and recover the predicted HR depth map to its original scale before loss computation, ensuring accurate depth prediction our MVS task. Additionally, we introduce a novel **confidence mask** for loss calculation to retain high-confidence LR estimates.

1) Multi-modal Feature Extraction: Our multi-modal feature extraction includes three branches to extract multi-level depth, cost, and guidance features $\left\{\mathbf{F}_d^j, \mathbf{F}_v^j, \mathbf{F}_g^j\right\}_{j=1}^{m}$ from LR depth, cost volume and HR guidance separately. Here, m, the number of layers for feature extraction, is set to 4, following the baseline [27]. The depth and guidance feature branches remain consistent with the baseline model, and an additional cost feature branch has been introduced. In this new branch, the proposed **CFE module** is first applied on the obtained cost and probability volumes $(\mathbf{V}_L, \mathbf{P}_L)$ to extract MVS-specific cost feature $\mathbf{F}_v^0$. Subsequently, the cost feature is processed using the same network structure as the other two branches. The procedures of the three feature extraction branches are detailed in Appendix A.

Here, we introduce the newly proposed **CFE module**, designed to encode the uncertainty in matching-based LR depths. It generates cost features $\mathbf{F}_v^0$, which are selectively activated in regions with high uncertainty. We begin by filtering out unreliable matching results from the cost volume $\mathbf{V}_L \in \mathbb{R}^{C\times D\times \frac{W}{4}\times \frac{H}{4}}$ using the probability volume $\mathbf{P}_L \in \mathbb{R}^{D\times \frac{W}{4}\times \frac{H}{4}}$. This step diminishes the scores of ambiguous matchings by assigning lower probability values, formally expressed as:

$$\mathbf{V}'_L = \sum_{c=1}^{C} \mathbf{P}_L^{d,w,h} \odot \mathbf{V}_L^{c,d,w,h}, \tag{1}$$

where $\odot$ denotes element-wise multiplication. The confidence map $\mathbf{U}_L \in \mathbb{R}^{\frac{W}{4}\times \frac{H}{4}}$, encoding depth uncertainty, is then derived by applying entropy along the depth dimension:

$$\mathbf{U}_L = f_u\bigg(-\frac{1}{\log(D)} \sum_{d=1}^{D} \mathbf{P}_L^{d,w,h} \log(\mathbf{P}_L^{d,w,h}) \bigg), \tag{2}$$

where $\log(D)$ normalizes the values to the range $(0, 1]$, and $f_u(\cdot)$ denotes a shallow 2D CNN to enhance the representation ability, following [20], [47]. Next, the cost feature $\mathbf{F}_v^0$ is extracted from the filtered cost volume $\mathbf{V}'_L$ and the confidence map $\mathbf{U}_L$ as follows:

$$\mathbf{F}_v^0 = Conv_1^0\big(|\mathbf{V}'_L, \mathbf{U}_L|_\tau\big), \tag{3}$$

where $|a, b|_\tau$ denotes a cut-off function that outputs a when b exceeds the threshold τ, and 0 otherwise. $Conv_1^0$ refers to a 1×1 convolutional layer that controls the cost feature channel number to be 32. The hyper-parameter τ, representing the confidence threshold, is set to 0.3 in this paper.

2) Uncertainty-aware Multi-Modal Fusion (UMF): After extracting multi-modal features, the key step in our uncertainty-aware GDSR is identifying and integrating relevant multi-modal information, incorporating uncertainty, to establish consistent structures for HR depth inference. Following the feature enhancement and fusion stages proposed in [27], depth, cost, and guidance features are systematically fused. First, reliable structures within each modality are emphasized to enhance depth, cost, and guidance features. Then, depth-uncertainty and depth-guidance features are separately identified to extract supportive signals, which are subsequently fused into a unified uncertainty-aware mutual structure representation.

Specifically, the feature enhancement stage employ gated convolution [48] to obtain enhanced depth, cost and guidance features ($\hat{\mathbf{F}}_d^j$, $\hat{\mathbf{F}}_v^j$, $\hat{\mathbf{F}}_g^j$), formulated as follows:

$$\begin{cases} \hat{\mathbf{F}}_d^j &= \sigma(Conv_{d,1}^j(\mathbf{F}_d^j)) \odot \phi(Conv_{d,2}^j(\mathbf{F}_d^j)), \\ \hat{\mathbf{F}}_v^j &= \sigma(Conv_{v,1}^j(\mathbf{F}_v^j)) \odot \phi(Conv_{v,2}^j(\mathbf{F}_v^j)), \\ \hat{\mathbf{F}}_g^j &= \sigma(Conv_{g,1}^j(\mathbf{F}_g^j)) \odot \phi(Conv_{g,2}^j(\mathbf{F}_g^j)), \end{cases} \tag{4}$$

where $Conv_{d,1}^j$, $Conv_{d,2}^j$; $Conv_{v,1}^j$, $Conv_{v,2}^j$; and $Conv_{g,1}^j$, $Conv_{g,2}^j$ are convolutional kernels for depth, cost and guidance, respectively, the subscripts 1 and 2 represent two different convolutional operations, and $\phi(\cdot)$ is the sigmoid function

to limit the output within the range of 0 and 1.

The feature fusion stage selectively emphasizes relevant signals from the depth and guidance streams before fusion, accounting for their distinct characteristics across depth, cost, and guidance modalities. First, depth-uncertainty and depth-guidance features are generated through concatenation, pooling, and pixel-wise summation. These features are then multiplied to obtain uncertainty-aware mutual structure features, denoted as $\big(\mathbf{F}^j_{e2}\big)$. Finally, the depth-uncertainty ($\mathbf{F}^j_{e1}$) and uncertainty-aware mutual structure ($\mathbf{F}^j_{e2}$) features are further encoded as excitation signals to guide the fusion of depth and guidance features. Formally, this process is expressed as:

$$\begin{cases} \mathbf{F}^j_{c1} = \sigma\Big(Conv^j_{f,1}\big([\hat{\mathbf{F}}^j_d, \hat{\mathbf{F}}^j_v]\big)\Big), \\ \mathbf{F}^j_{c2} = \sigma\Big(Conv^j_{f,2}\big([\hat{\mathbf{F}}^j_d, \hat{\mathbf{F}}^j_g]\big)\Big), \\ \mathbf{F}^j_{e1} = MaxPool\big(\mathbf{F}^j_{c1}\big) + VarPool\big(\mathbf{F}^j_{c1}\big), \\ \mathbf{F}^j_{e2} = \Big(AvgPool\big(\mathbf{F}^j_{c2}\big) + VarPool\big(\mathbf{F}^j_{c2}\big)\Big) \odot \mathbf{F}^j_{e1}, \\ \mathbf{E}^j_d = \phi\Big(Conv^j_{f,3}\big([\mathbf{F}^j_{e1}, \mathbf{F}^j_{e2}]\big)\Big), \\ \mathbf{E}^j_g = \phi\Big(Conv^j_{f,4}\Big(\mathbf{F}^j_{e2}\Big)\Big), \\ \mathbf{F}^j_f = \mathbf{E}^j_d \odot \hat{\mathbf{F}}^j_d + \mathbf{E}^j_g \odot \hat{\mathbf{F}}^j_g, \end{cases} \tag{5}$$

where $[\cdot,\cdot]$ denotes concatenation operation, $Conv^j_{f,n}$ denotes convolutional operations, $\mathbf{E}^j_g$ and $\mathbf{E}^j_d$ are guidance and depth features excitation signals, respectively, $MaxPool(\cdot)$, $AvgPool(\cdot)$ and $VarPool(\cdot)$ indicate the maximal, average and variance pooling operations, respectively. $\mathbf{F}^j_f$ represents the final fused multi-modal feature. As will be shown later in Fig. 8 of Subsection IV-D *Ablation Study*, unlike the baseline whose fused features exhibit error-copying artifacts, our uncertainty-aware GDSR effectively mitigates such issues. This underscores the effectiveness of the CFE and UMF modules in accurately perceiving and handling the uncertainty inherent in LR depth (encoded by the cost features), thereby facilitating the exploration of correct mutual structures across RGB-D modalities to achieve reliable HR depth prediction.

3) Supervision of Uncertainty-aware GDSR: After the multi-modal feature fusion step, we employ the bi-directional hierarchical feature collaboration (BHFC) and final HR depth reconstruction (RC) blocks of AMHF [27] to output HR depth $\mathbf{D}_{SR}$, formulated as follows:

$$\mathbf{D}'_{SR} = RC\Big(\big\{BHFC_j\big(\mathbf{F}^j_d, \mathbf{F}^j_c, \mathbf{F}^j_g\big)\big\}^4_{j=1}\Big), \tag{6}$$

$$\mathbf{D}_{SR} = \mathbf{D}'_{SR} + \mathbf{D}^{\uparrow}_L, \tag{7}$$

where $\mathbf{D}^{\uparrow}_L$ denotes the bicubic upsampled LR depth map. The applied Super-Resolution (SR) loss and confidence mask are described as follows:

$$\mathcal{L}_{SR} = \frac{1}{|\Omega|}\sum_{\Omega}\big\|\mathbf{D}^{gt}_{HR} - \mathbf{D}_{SR}\big\|_{S1} + \frac{1}{|\Omega|}\sum_{\Omega}\mathbf{M}_c\|\mathbf{D}'_{SR}\|_1, \tag{8}$$

where Ω is the ground-truth reference depth valid region, $\mathbf{D}^{gt}_{HR}$ is the ground-truth reference depth, $\|\cdot\|_1$ denotes the L1 norm, and $\|\cdot\|_{S1}$ denotes the smoothed L1 norm, while $\mathbf{M}_c$ denotes the proposed **confidence mask**.

The first term in Eq. (8) encourages global accuracy in the depth super-resolution process. Different from AMHF that formulates the loss function using the L1 distance in the normalized space, we rescale the output HR depth to the original depth range and replace the L1 distance with the smooth L1 distance, to promote accurate depth value prediction in our MVS scenario.

The second term, a residual loss, is devised to preserve reliable matching-based depth estimates and enable large corrections in uncertain regions. To supervise the residual depth $\mathbf{D}'_{SR}$, we introduce a confidence mask $\mathbf{M}_c$, derived from the upsampled LR confidence map $\mathbf{U}^{\uparrow}_L$ as follows:

$$\mathbf{M}_c(\boldsymbol{p}) = \begin{cases} \exp\Big(-\big(\frac{(\mathbf{U}^{\uparrow}_L(\boldsymbol{p})-\mu)^2}{2\sigma^2_{\mathrm{m}}}\big)\Big), & \mathbf{U}^{\uparrow}_L(\boldsymbol{p}) < \epsilon, \\ 1, & \mathbf{U}^{\uparrow}_L(\boldsymbol{p}) \geq \epsilon, \end{cases} \tag{9}$$

where $\mathbf{U}^{\uparrow}_L(\boldsymbol{p})$ denotes the depth confidence of pixel $\boldsymbol{p}$, derived from Eq. (2) through the entropy operation along the depth dimension, while the parameter ϵ signifies the confidence threshold. When the confidence of a pixel exceeds ϵ, indicating high confidence, we assign the value 1 to the corresponding pixel in the confidence mask. In these highly confident regions, the residual depth is encouraged to remain small to preserve the matching-based LR estimates. Conversely, for pixels falling below the confidence threshold, a soft constraint is applied via the pre-defined Gaussian distribution of mean μ and standard deviation σ_{m}, to reduce the influence of their residuals (see Fig. 11 of Appendix A for an illustration of Eq. (9)). In ambiguous regions, the first term in Eq. (8) becomes dominant, allowing the model to predict larger residual corrections. To cautiously constrain depth residuals only for highly reliable regions, hyper-parameters μ, ϵ, and σ_{m} are set to 0.9, 0.9, and 0.049, respectively.

An example of resulting confidence mask will be shown in Fig. 7(b) of Subsection IV-D *Ablation Study*. It aligns closely with the upsampled LR error map in Fig. 7(i), indicating that our confidence mask can accurately identify reliable point depths. In these regions, residuals are constrained to remain small. Conversely, in less reliable areas, the first term in Eq. (8) dominates, guiding RGB-D structure-consistent HR depth prediction. A comparison between Fig. 7(j) and Fig. 7(k) demonstrates a noticeable reduction in depth estimation errors when incorporating the proposed confidence mask.

C. Matching-based Depth Refinement

After the uncertainty-aware GDSR step, we propose refining the obtained depth map $\mathbf{D}_{SR}$ by incorporating critical multi-view feature matching consistency, which is essential for our MVS task. Inspired by learnable PatchmatchNet [12], which utilizes deformable convolutions to find related neighbor depths, our refinement involves depth self-reintegration with learnable neighbor depths, weighted by their aggregated feature similarity scores. Our intention is to evaluate the reliability of current depth estimation through multi-view observations and then reintegrate related neighbors to the final refined depth, weighted by their reliabilities. Our refinement process is more lightweight than the cost volume one because the neighbor weights are computed based on their current estimation $\mathbf{D}_{SR}$, allowing evaluation with a single depth candidate

rather than a set of candidates. For each pixel, we first sampled K_e related neighbor locations, which is determined by adding two components: (1) predefined grid offsets within the local spatial window, $\{\boldsymbol{p}_k\}_{k=1}^{K_e}$, and (2) per-pixel adaptive offsets, $\{\triangle\boldsymbol{p}_k\}_{k=1}^{K_e}$, derived from the reference feature map $\mathbf{F}_0$. Then, the refined depth $\mathbf{D}(\boldsymbol{p})$ and the corresponding confidence map $\mathbf{U}(\boldsymbol{p})$ are defined as follows:

$$\begin{aligned}\mathbf{D}(\boldsymbol{p}) =& \frac{1}{\sum_{k=1}^{K_e} m_k}\sum_{k=1}^{K_e} m\big(\boldsymbol{p}+\boldsymbol{p}_k+\triangle\boldsymbol{p}_k\big)\\ &\times \mathbf{D}_{SR}\big(\boldsymbol{p}+\boldsymbol{p}_k+\triangle\boldsymbol{p}_k\big),\end{aligned} \tag{10}$$

$$\mathbf{U}(\boldsymbol{p}) = \sum_{k=1}^{K_e} m(\boldsymbol{p}+\boldsymbol{p}_k+\triangle\boldsymbol{p}_k), \tag{11}$$

where K_e is the hyper-parameter defining the number of sampled neighbors, which is set to 12, and m_k is the weight factor of the k-th neighbor, calculated based on their aggregated similarity scores.

The computation of neighbor similarity scores involves common operations of the cost volume pipeline, including differentiable warping, group-wise correlation and multi-view cost aggregation operations. This process is formulated as follows:

$$\boldsymbol{p}' = \mathbf{K}_i\mathbf{T}\big(\mathbf{K}_0^{-1}\mathbf{D}_{SR}(\boldsymbol{p})\big), \tag{12}$$

$$\mathbf{S}_i(\boldsymbol{p})^g = \frac{G}{C}\left\langle \mathbf{F}_0(\boldsymbol{p})^g, \mathbf{F}_i(\boldsymbol{p}')^g\right\rangle, \tag{13}$$

$$\mathbf{S}(\boldsymbol{p})^g = \frac{\sum_{i=1}^{N-1}\mathbf{W}_i(\boldsymbol{p})\mathbf{S}_i(\boldsymbol{p})^g}{\sum_{i=1}^{N-1}\mathbf{W}_i(\boldsymbol{p})}, \tag{14}$$

$$m(\boldsymbol{p}) = \phi\big(Conv_1(\mathbf{S}(\boldsymbol{p})^g)\big). \tag{15}$$

Given the reference and source feature maps, $\mathbf{F}_0$, $\{\mathbf{F}_i\}_{i=1}^N$, and the intrinsic parameters, $\mathbf{K}_0$, $\{\mathbf{K}_i\}_{i=1}^N$, the extrinsic transformation from the reference to the source view is denoted as $\mathbf{T}$, based on the current estimation $\mathbf{D}_{SR}$, and the pixel coordinates $\boldsymbol{p}'$ in source view i corresponding to the pixel $\boldsymbol{p}$ in the reference view can be calculated via Eq. (12). The warped source feature is obtained through bilinear interpolation, denoted as $\mathbf{F}_i(\boldsymbol{p}')$. Then, we organize the channels of the feature maps into G groups along the channel dimension, resulting in $\mathbf{F}_0(\boldsymbol{p})^g$ and $\mathbf{F}_i(\boldsymbol{p}')^g$. The similarity within the g-th group, denoted as $\mathbf{S}_i(\boldsymbol{p})^g \in \mathbb{R}^{W\times H\times G}$, is computed using (Eq. 13), where $\langle\cdot,\cdot\rangle$ denotes the inner product. Once the two-view feature similarity scores are calculated for each group, pixel-wise view weights $\{\mathbf{W}_i\}_{i=1}^{N-1}$, $\mathbf{W}_i \in \mathbb{R}^{\frac{W}{4}\times\frac{H}{4}}$, obtained during the LR depth estimation step, are upsampled to the consistent spatial resolution for final per-group similarities calculation, resulting in $\mathbf{S}(\boldsymbol{p})^g \in \mathbb{R}^{W\times H\times G}$. Subsequently, a small network with 2D 1×1 convolution ($Conv_1$) and sigmoid non-linearity function (ϕ) is applied to output per-pixel weight value $m(\boldsymbol{p})$.

D. Loss Function

To optimize the estimated depth, we employ three components of depth loss dedicated to supervising the outputs of the three distinct steps:

$$\mathcal{L}_{Total} = \mathcal{L}_{LR} + \mathcal{L}_{SR} + \mathcal{L}_{Ref}. \tag{16}$$

$\mathcal{L}_{LR}$ defines the Cross Entropy (CE) loss that applies on the iteratively produced probability volume $\mathbf{P}_L$ and the ground-truth one-hot occupancy volume $\mathbf{G}$ for valid pixels [11], defined as follows:

$$\mathcal{L}_{LR} = \sum_{it=1}^{4}\frac{1}{|\Omega|}\sum_{\boldsymbol{p}\in\Omega}\sum_{d=1}^{D} -\mathbf{G}^{it}(d,\boldsymbol{p})\log\big(\mathbf{P}_L^{it}(d,\boldsymbol{p})\big), \tag{17}$$

where Ω denotes the set of valid pixels, and $|\Omega|$ represents the count of valid pixels.

$\mathcal{L}_{SR}$, defined in Eq. (8), serves to supervise the output of the GDSR step.

$\mathcal{L}_{Ref}$ represents the uncertainty-aware loss function applied to the refined depth, following [20], [47], [49]:

$$\begin{aligned}\mathcal{L}_{Ref} =& \frac{1}{|\Omega|}\sum_{\boldsymbol{p}\in\Omega}\frac{\big|\mathbf{D}_{HR}^{gt}(\boldsymbol{p})-\mathbf{D}(\boldsymbol{p})\big|}{\mathbf{U}(\boldsymbol{p})} + \log(\mathbf{U}(\boldsymbol{p}))\\ =& \frac{1}{|\Omega|}\sum_{\boldsymbol{p}\in\Omega}\frac{\big|\mathbf{D}_{HR}^{gt}(\boldsymbol{p})-\mathbf{D}(\boldsymbol{p})\big|}{\exp\big(-\log(\mathbf{U}(\boldsymbol{p}))\big)} + \log(\mathbf{U}(\boldsymbol{p})),\end{aligned} \tag{18}$$

where $U(\boldsymbol{p})$ denotes the predicted confidence map. The first term encourages the network to produce lower confidence for pixels exhibiting higher biases and vice versa. The second regularization term is designed to prevent the model from outputting uniformly low confidence for all pixels. This dual-term formulation aids in implicitly learning uncertainty during the training procedure.

E. Iterative Depth Fusion for 3D Reconstruction

Estimated multi-view depth maps often exhibit biased estimates in challenging regions such as the sky or textureless surfaces. To address this issue, a common approach is to fuse high-confidence and multi-view consistent depth points to generate the final point cloud. Current methods typically apply one-time photometric and geometric consistency filtering to each depth map. However, while this effectively eliminates unimportant background areas, it can also result in holes in textureless regions, leading to incomplete reconstructions. To improve this process, we propose to apply iterative depth self-reintegration and geometric consistency filtering until the number of valid points stabilizes. Initially, we perform one-time photometric filtering by applying confidence map $\mathbf{U}(\boldsymbol{p}) \le \xi_c$, as in the previous works [3], [11], [14]. Next, we alternate between geometric consistency filtering and a simplified, non-parametric depth self-reintegration process.

We begin with the geometric consistency filtering. For each estimated reference depth $\mathbf{D}_0(\boldsymbol{p})$, each pixel $\boldsymbol{p}$ in the reference image $\mathbf{I}_0$ is back-projected into 3D space using camera parameters and the estimated depth, yielding a 3D point $\boldsymbol{P}$. This point is then projected onto neighboring images $\mathbf{I}_i$ to generate the corresponding pixel $\boldsymbol{q}$, and the process is reversed by back-projecting the pixel $\boldsymbol{q}$ from the neighbor view with the estimated depth $\mathbf{D}_i(\boldsymbol{q})$ into 3D space and reprojecting it back to the reference image, denoted as $\boldsymbol{p}'$. Based on these operations, the 3D point $\boldsymbol{P}$ is considered consistent in the neighbor image $\mathbf{I}_i$ if it satisfies the following pixel and depth

reprojection error thresholds:

$$\|\boldsymbol{p} - \boldsymbol{p}'\|_2 \leq \xi_p, \tag{19}$$

$$\frac{\|\mathbf{D}_0(\boldsymbol{p}) - \mathbf{D}_0(\boldsymbol{p}')\|}{\mathbf{D}_0(\boldsymbol{p})} \leq \xi_d, \tag{20}$$

where ξ_p and ξ_d are the pixel and depth reprojection error thresholds, respectively. If at least ξ_N neighboring images satisfy these thresholds, the depth of pixel $\boldsymbol{p}$ is deemed multi-view consistent and is preserved. This results in a cross-validated mask defining geometrically consistent pixels.

Subsequently, with the cross-validated mask and filtered depth map, a non-parametric depth self-reintegration procedure is applied to fill cross-invalidated pixels, similar to the earlier refinement step. The neighbor pixel positions are accessed by offsets $\{\triangle \boldsymbol{p}_k\}_{k=1}^{K_e}$ of Eq. (10). However, in this step, neighbor depth values are reintegrated by averaging, without using separate weights from feature matching scores. Meanwhile, the valid mask is applied to reintegrate only the depths that have been cross-validated.

After each iteration, we track the growth rate of the validated pixel count in the filtered depth map. If the growth rate R between the current and previous iterations is less than $\xi_R = 0.01$, we consider the contribution of this reference depth map to the point cloud as stabilized. Then, we stop iterating and process the next estimated reference depth map. The calculation of the growth rate is formulated as follows:

$$R = \frac{|N_c - N_p|}{N_p}, \tag{21}$$

where N_c is the validated pixel count in the current iteration, and N_p is the validated pixel count from the previous iteration.

TABLE I: Point cloud quantitative results on DTU [7]

Methods	Year	Mean Error Distance (mm)			Mem. (MB)	Run-Time (s)
		Acc.	Comp.	Overall		
Gipuma [50]	2015	**0.283**	0.873	0.578	-	-
Colmap [1]	2016	0.400	0.664	0.532	-	-
MVSNet [3]	2018	0.396	0.527	0.462	10823	1.21
UCSNet [14]	2020	0.338	0.349	0.344	4057	**0.37**
CasMVSNet [15]	2020	0.325	0.385	0.355	4591	0.49
PatchmatchNet [12]	2020	0.427	0.277	0.352	<u>1629</u>	0.51
PVA-MVSNet [51]	2020	0.379	0.336	0.357	25466	1.01
Vis-MVSNet [47]	2020	0.369	0.361	0.365	4775	0.58
CDS-MVSNet [17]	2022	0.352	0.280	0.316	4492	0.66
NP-CVP-MVSNet [29]	2022	0.356	0.275	0.315	6054	1.20
GBi-Net$^\diamond$ [11]	2022	0.312	0.293	0.303	2108	0.61
Prior-Net [49]	2023	0.351	0.287	0.319	8397	0.64
UniMVSNet [16]	2022	0.352	0.278	0.315	6139	0.83
UGNet [20]	2022	0.334	0.330	0.332	-	-
NR-MVSNet [34]	2023	0.331	0.285	0.308	5649	0.57
PFR-MVSNet [52]	2023	0.289	0.383	0.336	15063	3.76
ARAI-MVSNet [35]	2023	0.292	0.334	0.313	5386	0.61
MVSFormer [18]	2023	0.327	0.251	<u>0.289</u>	4970	0.48
WT-MVSNet [19]	2023	0.309	0.281	0.295	5221	0.79
DMVSNet [21]	2023	0.338	0.272	0.305	6672	0.88
MVSTR [30]	2023	0.356	0.295	0.326	3879	0.82
HAMMER [22]	2024	0.326	0.270	0.298	3175	-
GoMVS [32]	2024	0.347	0.227	**0.287**	7901	0.70
ICV-Net [13]	2025	<u>0.286</u>	0.347	0.317	**1221**	0.59
SR-MVSNet* (Ours)	-	0.372	**0.212**	0.292	2839	<u>0.46</u>
SR-MVSNet (Ours)	-	0.387	<u>0.215</u>	0.301	2839	<u>0.46</u>

Lower is better for Accuracy (Acc.), Completeness (Comp.), and Overall. The best results are in **Bold** and the second best are in <u>underlined</u>. $\diamond$ denotes that GBi-Net is re-tested with same post-processing threshold to all scans for fair comparisons with other methods.

IV. Experiments

A. Datasets and Evaluation Metrics

Datasets: We employ the following four datasets for training and evaluation, namely, 1) **DTU** [7], which consists of 124 scenes captured in laboratory settings, under seven lighting conditions, 2) **Tanks & Temples (TnT)** [8], which contains more complex and realistic 14 scenes captured in real environments, provided as a set of video sequences with image resolutions about 1920×1080. 3) **BlendedMVS** [10], a synthetic dataset, which contains 17k images of 113 diverse scenes, and 4) **ETH3D** [9], which contains 13 scenes in training set and 12 scenes in test set with HR images of 6212×4140. Following previous works [12], [18], [34], we regard the training set of ETH3D as one of the test sets.

Evaluation Metrics: For the DTU benchmark, we adopt the mean accuracy (Acc.) and completeness (Comp.) of the distance metrics for point cloud evaluation. The Overall metric takes the average of Acc. and Comp., providing a comprehensive measure of reconstruction quality. Notably, there is sometimes an inherent trade-off between accuracy and completeness. In our analysis, the Overall is the most critical metric for reconstruction performance evaluation. Besides, we adopt the mean absolute depth error (MAE), the area under sparsification error curve (AUSE), and precision metrics (e_2, e_4, e_8) for the depth map and confidence map evaluation. The precision metric e_α is defined as the pixel percentage within α mm error, the bigger the better. For the TnT and ETH3D benchmarks, we adopt the F-score and F_1-score metrics for point cloud evaluation. they are the harmonic mean of points percentages with precision and recall at 2 mm distance threshold. The evaluation procedure is conducted online through official platforms after submitting our reconstruction results.

B. Implementation Details

Our models were initially trained on the DTU [7] training set and evaluated on the DTU testing set. Then, the model is fine-tuned on BlendedMVS dataset [10] for generalizability evaluation on TnT [8] and ETH3D [9] benchmarks.

Following common MVS training protocols [3], [12], [14], [16], [17], [25], [34], [35], we trained "SR-MVSNet" using half-resolution DTU data (640 × 512) with 5 input images. Fine-tuning on BlendedMVS employed a resolution of 768 × 576 and 7 input images. The network was trained for 16 epochs [11], [15], [19], [25], [34] with the Adam optimizer, starting at a learning rate of 0.0001, which was halved after epochs 10, 12, and 14. The batch size was 4, and training was conducted on two NVIDIA RTX 3090 GPUs. In line with recent advancements [11], [18], [19], [21], [22], we train the model "SR-MVSNet*" with full-resolution DTU data (1600×1200) and employ the random cropping pre-processing step as GBi-Net [11], while other training setups followed common protocols. For TnT evaluation, we expanded the candidate source views similar to MVSFormer [18], enhancing performance in complex scenes. Both models applied the proposed iterative fusion post-processing strategy, using consistent hyper-parameters throughout the dataset.

TABLE II: Quantitative results of F-score on Tanks & Temples dataset [8]

Method	Year	Intermediate									Advanced						
		Mean	Fam.	Fra.	Hor.	Lig.	M60.	Pan.	Pla.	Tra.	**Mean**	Aud.	Bal.	Cou.	Mus.	Pal.	Tem.
COLMAP [1]	2016	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
UCSNet [14]	2020	54.83	76.09	53.16	43.03	54.00	55.60	51.49	57.38	47.89	-	-	-	-	-	-	-
CasMVSNet [15]	2020	56.42	76.36	58.45	46.20	55.53	56.11	54.02	58.17	46.56	31.12	19.81	38.46	29.10	43.87	27.36	28.11
PatchmatchNet [12]	2020	53.15	66.99	52.64	43.24	54.87	52.87	49.54	54.21	50.81	32.31	23.69	37.73	30.04	41.80	28.31	32.29
CDS-MVSNet [17]	2022	60.82	78.17	61.74	53.12	60.25	61.91	58.45	**62.35**	50.58	-	-	-	-	-	-	-
NP-CVP-MVS [29]	2022	59.64	78.93	64.09	51.82	59.42	58.39	55.71	56.07	52.71	-	-	-	-	-	-	-
GBi-Net [11]	2022	61.42	79.77	67.69	51.81	61.25	60.37	55.87	60.67	53.89	37.32	29.77	42.12	36.30	47.69	31.11	36.93
Prior-Net [49]	2022	60.63	79.02	60.93	51.65	60.52	61.78	56.19	61.37	53.59	34.61	26.99	40.46	30.76	47.81	28.96	32.71
UniMVSNet [16]	2022	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
UGNet [20]	2022	63.12	79.61	63.35	50.32	66.36	64.80	60.84	62.25	57.41	37.12	23.28	43.49	36.04	50.59	31.81	37.54
NR-MVSNet [34]	2023	62.94	80.78	63.55	53.09	60.61	65.29	62.20	60.65	57.31	37.20	26.76	43.21	35.79	50.01	33.35	34.08
PFR-MVSNet [52]	2023	64.56	81.57	65.50	54.43	63.37	**66.44**	**65.65**	62.31	57.23	39.22	28.73	44.97	39.83	53.46	33.79	34.52
ARAI-MVSNet [35]	2023	61.09	79.48	66.83	54.15	59.56	58.58	57.38	56.51	56.27	38.68	26.13	43.01	38.63	48.88	35.39	40.01
MVSFormer [18]	2023	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
WT-MVSNet [19]	2023	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
DMVSNet [21]	2023	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
MVSTR [30]	2023	56.93	76.92	59.82	50.16	56.73	56.53	51.22	56.58	47.48	32.85	22.83	39.04	33.87	45.46	27.95	27.97
HAMMER [22]	2024	61.70	78.45	59.25	54.33	62.80	63.20	59.57	61.72	54.23	36.13	24.17	40.07	38.14	49.56	31.54	33.31
GoMVS [32]	2024	**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**
ICV-Net [13]	2025	55.56	72.05	56.87	41.27	52.32	58.66	53.59	59.10	50.64	34.70	26.01	41.53	33.70	41.98	29.34	35.63
SR-MVSNet*	-	65.36	82.55	**69.52**	57.69	65.11	64.96	62.66	59.56	60.78	41.48	24.48	**47.87**	**42.86**	52.94	36.40	44.34
SR-MVSNet$^*_{dyc}$	-	64.84	81.26	69.37	57.83	64.23	63.93	62.61	58.88	60.59	41.18	25.07	47.46	41.57	52.21	**36.55**	44.24

We report the F-score metric. **"Mean"** refers to the average F-score of all scenes. The best results are in **Bold** and the second best results are in underlined. All the values, including ours, are available in the website [53].

TABLE III: Results on Tanks & Temples dataset [8] using the same fusion parameter setting across all scenes.

Method	Year	**Mean**	Fam.	Fra.	Hor.	Lig.	M60.	Pan.	Pla.	Tra.	**Mean**	Aud.	Bal.	Cou.	Mus.	Pal.	Tem.
GBi-Net [11]	2022	60.32	79.29	65.07	49.35	60.41	59.79	55.30	59.52	53.80	33.93	22.69	37.30	32.96	46.37	29.23	35.03
UniMVSNet [16]	2022	60.39	79.28	65.59	41.67	63.71	61.58	58.77	60.33	52.18	28.05	13.77	29.95	26.70	46.12	28.40	23.36
MVSFormer [18]	2023	63.41	80.62	65.80	54.35	64.73	**64.47**	**61.88**	56.94	58.51	36.67	22.99	42.38	34.93	49.22	33.17	37.33
HAMMER [22]	2024	61.70	78.45	59.25	54.33	62.80	63.20	59.57	**61.72**	54.23	36.13	24.17	40.07	**38.14**	49.56	31.54	33.31
GoMVS [32]	2024	62.16	**81.48**	**69.21**	45.97	64.09	63.35	56.39	55.83	**60.92**	36.37	20.52	39.23	31.62	49.20	34.36	**43.27**
SR-MVSNet	-	**64.06**	80.68	67.23	**57.94**	**66.27**	62.29	60.58	60.58	57.19	**39.52**	**25.63**	**46.12**	37.57	**50.95**	**35.25**	41.60
SR-MVSNet$_{dyc}$	-	63.69	80.52	67.08	57.92	65.43	62.14	60.53	58.35	57.56	39.19	24.56	46.02	37.50	50.77	35.19	41.10

GoMVS [32] is re-tested using the average parameter configuration recommended by the authors. MVSFormer [18] results are derived by retraining on the mainstream half-resolution DTU data and using consistent candidate source views with other methods. Results for GBi-Net [11], UniMVSNet [16] and HAMMER [22] are from the paper [22].

C. Comparisons with State-of-the-art Methods

Results on DTU Dataset: We evaluate our proposed method on the DTU dataset with a fixed testing resolution of 1600×1152 and using five views ($N = 5$). Our evaluation comprises two protocols: firstly, using the official DTU evaluation toolbox [7] to compare reconstructed point clouds against ground-truth 3D scans; secondly, evaluating depth map accuracy using MAE and precision metrics, as detailed in Subsection Appendix B.

Quantitative results of reconstruction evaluation are shown in Table I, and qualitative ones are shown in Fig. 12 of the Appendix. Our best-performing model, SR-MVSNet*, ranked third in the overall metric. Compared to the first-place method, GoMVS [32], which incorporates geometric-consistent propagation during the cost aggregation step, our model achieves higher completeness but lower accuracy. GoMVS's well-designed geometric-consistent cost aggregation is effective in both generally textured regions and local weakly textured areas. Moreover, its use of denser depth sampling across multiple stages enables more precise depth estimates, which is particularly advantageous in small-scale DTU scenes. However, this comes at the cost of efficiency: GoMVS's cascaded cost volume formulation results in approximately 2.8 × higher GPU memory consumption and significantly longer inference time compared to our approach. In comparison to the second-place method, MVSFormer [18], a recent transformer-based model also built upon the cascaded cost volume framework, our model achieves competitive reconstruction quality while reducing memory usage by 42%. Notably, our method obtains the highest completeness score, suggesting that the proposed RGB-D structural consistency learning effectively improves the robustness of depth estimation, particularly in handling scenes with local or large areas of ambiguous matching. When evaluated against lightweight and iterative MVS approaches such as PatchmatchNet [12] and GBi-Net [11] (used for our LR depth estimation), our model demonstrates superior overall performance with reduced inference time, while with slight increase in memory usage (731 MB). Overall, these results illustrate that our approach effectively integrates the complementary strengths of cost volume and depth super-resolution techniques to achieve advanced reconstruction quality while decreasing memory footprint and runtime as much as possible. This represents a novel and efficient alternative to conventional cascaded cost volume architecture for MVS reconstruction.

Results on Tanks & Temples Dataset: We evaluate the generalization capability of our approach using the TnT benchmark, with $N = 11$ input images, and the input image sizes are 1920×1024 or 2048×1024 to make the images divisible by 64. Table II compares our method with state-of-the-art MVS methods, noting that these methods are trained with

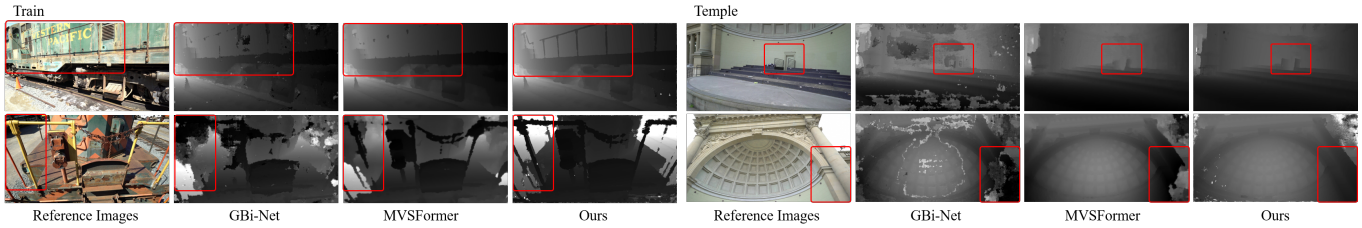


Fig. 4: Depth estimation of GBi-Net [11], MVSFormer [18] and our method on Tanks & Temples benchmark.

different data and employ distinct post-processing methods, as summarized in Table VIII of Subsection Appendix A.

Our best-performing model, SR-MVSNet*, achieves the third-highest F-score on the intermediate set and the second-highest F-score on the advanced set of the benchmark. In comparison to high-performing methods such as GoMVS [32] and MVSFormer [18], both methods require the time-consuming manual search for optimal fusion parameters for each scene. In contrast, our proposed post-processing strategy does not rely on per-scene hyperparameter adjustment, which, while more practical, is challenging to achieve the highest F-score in every scenario. Additionally, our uncertainty-aware GDSR module may face challenges in indoor and outdoor scenes with complex lighting conditions, such as point light sources, strong shadows, or reflections. These factors may limit our uncertainty-aware GDSR module's ability to capture real depth-discontinuous structural information crucial for accurate HR depth estimation. In such cases, setting stricter post-processing thresholds could filter out depth error but require tuning fusion parameters as done in some previous works [11], [18], [21], [32]. Furthermore, MVSFormer [18] necessitates multiple data augmentation steps and elaborate multi-scale training processes to achieve optimal results. GoMVS [32] needs to perform time-consuming preprocessing of monocular normal estimation in a divide-and-conquer manner. In contrast, our model, SR-MVSNet*, delivers competitive reconstruction performance without these additional complexities. More importantly, our method significantly decreased GPU memory consumption and inference time compared to other high-performing methods, including MVSFormer [18], WT-MVSNet [19], DMVSNet [21] and GoMVS [32], as demonstrated in the DTU evaluation. Qualitative comparisons in Fig. 4 show that our HR depth estimations excel in capturing detailed structures by leveraging both multi-view feature matching and the structural consistency of RGB-D data. In contrast, GBi-Net [11] and MVSFormer [18], which rely solely on feature matching in cascaded cost volumes, tend to reach locally optimal solutions in certain areas and struggle in large-scale scenes due to insufficient depth sampling. GBi-Net, in particular, produces significant errors in textureless regions due to its limited use of only four depth candidates per iteration.

Table III further compares our mainstream-trained model, SR-MVSNet, with several high-performing methods on the benchmark, using consistent training data and invariant hyper-parameter settings across the entire dataset. Notably, under this more rigorous and fair experimental setup, our method demonstrates clear performance advantages over the listed approaches, including recent state-of-the-art models such as MVSFormer [18] and GoMVS [32]. When adopting the common dynamic fusion post-processing strategy [36] (Table II, Table III), both models, SR-MVSNet$_{dyc}$ and SR-MVSNet$^{*}_{dyc}$, still achieved good performance and show advantage on the TnT Advanced set, which includes more complex indoor scenes, exhibits greater variation in depth ranges for each reference view compared to the Intermediate set. These results underscore the effectiveness of our primary contribution: the "cost volume + depth super-resolution" framework, particularly in addressing wide-baseline, large-scale scenes with diverse depth ranges for each reference view, regardless of the post-processing strategy employed.

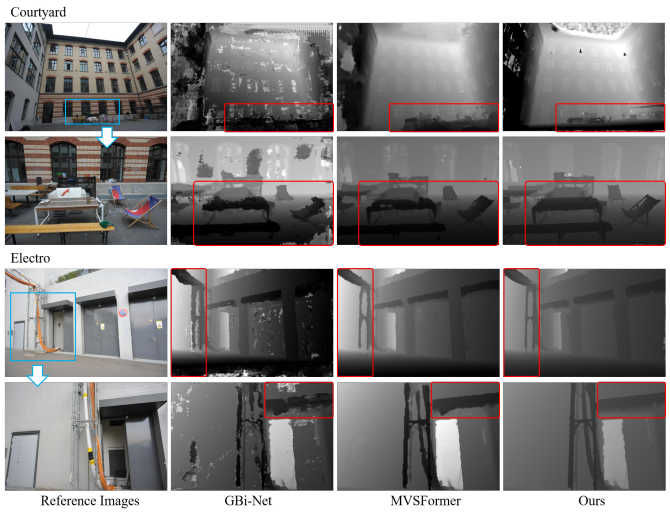


Fig. 5: Depth prediction of GBi-Net [11], MVSFormer [18] and our method on ETH3D benchmark.

Results on ETH3D Test Set: We further evaluate our SR-MVSNet* model on the ETH3D test set, with image resolution 1920×1280 and input view number $N = 7$. Quantitative comparisons with state-of-the-art learning-based MVS methods are presented in Table IV. Our method achieves the highest F_1-score across most scenes, demonstrating its strong generalization capability in handling wide-baseline, large-scale scenes. Fig. 5 shows depth predictions from sampled reference views generated by the baseline GBi-Net, MVSFormer [11], [18] and our method. The baseline GBi-Net [11] produces noticeable errors in textureless areas, where its iterative binary search strategy struggles in large-scale scenes with wide and varied depth ranges. This increases the risk of converging to local optima, leading to inaccurate depth estimation. In comparison, our method effectively corrects these outliers

TABLE IV: Quantitative results of F_1-score on ETH3D test set [9].

Method	Year	**Mean**	Bot.	Bou.	Bri.	Door	Exh.	Lec.	Liv.	Lou.	Obs.	Old.	Sta.	Ter.
PatchmatchNet [12]	2020	73.12	83.18	60.85	79.63	78.57	64.13	71.73	79.81	51.2	85.97	57.4	76.36	88.66
CDS-MVSNet [17]	2022	79.07	87.6	68.23	84.4	86.2	67.49	76.36	86.54	61.54	90.79	61.86	87.33	90.49
GBi-Net [11]	2022	78.4	87.6	63.01	88.39	89.28	71.63	75.34	87.28	54.8	91.11	62.72	79.57	90.09
UniMVSNet [16]	2022	81.6	80.06	64.41	87.6	**93.02**	78.42	78.56	88.16	69.78	93.26	69.58	76.49	91.6
NR-MVSNet [34]	2023	80.23	87.8	63.34	86.31	91.09	**79.01**	72.17	89.07	67.2	92.81	66.08	75.65	92.2
MVSFormer [18]	2023	82.85	87.35	70.21	90.5	90.38	68.55	81.32	89.47	61.68	93.46	78.98	90.26	92.05
GoMVS [32]	2024	85.91	89.84	69.44	90.45	93.00	76.70	**88.01**	91.23	**73.28**	93.91	81.50	88.80	**94.77**
SR-MVSNet*	-	**86.73**	**94.11**	**72.11**	**92.37**	92.56	74.50	86.84	**94.65**	67.31	**95.77**	**84.80**	**92.85**	92.84
SR-MVSNet$^*_{dyc}$	-	85.77	93.06	70.63	91.88	92.27	73.53	85.69	94.45	65.00	95.56	82.75	91.87	92.50

The evaluation metric is the F_1-score using percentage metric, which considers both accuracy and completeness of final reconstructed point cloud results. Higher F_1-score means a better reconstruction quality. All the values, including ours, are available in the website [54].

while remaining robust to variations in depth range across reference views, producing RGB-D structurally consistent HR reconstructions. Additionally, multi-view feature matching is performed both before and after the depth super-resolution step to ensure procedure accuracy. Ultimately, our method demonstrates superior capability in recovering depth for views with large and varying depth ranges, outperforming both the baseline GBi-Net [11] and recent state-of-the-art methods, including MVSFormer [18] and GoMVS [32]. When employing the common dynamic fusion post-processing strategy [36], our variant SR-MVSNet$^*_{dyc}$ still surpasses MVSFormer by a significant margin and achieves comparable performance to GoMVS, while requiring substantially less GPU memory. The ETH3D test set, characterized by its broad depth range distribution, particularly highlights the strength of our approach. By effectively exploring uncertainty-aware mutual structural cues from RGB-D inputs, our method shows clear advantages over cascaded MVS architectures in terms of both robustness and efficiency under challenging conditions. Apart from that, our iterative depth fusion post-processing strategy further enhances reconstruction performance across all benchmark evaluations.

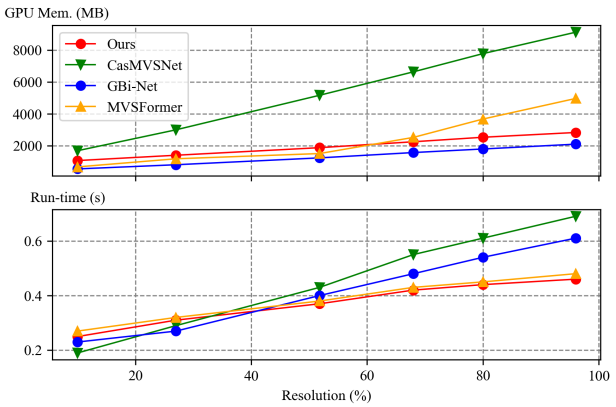


Fig. 6: GPU memory and run-time vs. input resolution on the DTU evaluation set [7]. The original resolution is 1600 × 1200 (100%), the highest tested resolution is 1600 × 1152 (96%).

Run-time and Memory Analysis: In this section, we evaluate the memory consumption and run-time of the proposed method. Since the depth super-resolution process constitutes the primary computational overhead in our framework, we compare our "cost volume + depth super-resolution" approach

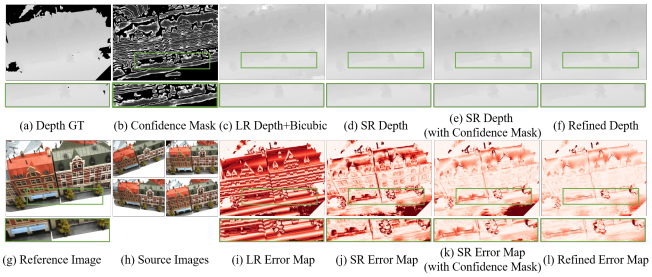


Fig. 7: Illustration of each component's influence on a reference depth estimation. (b) presents the produced confidence mask using Eq. (9), and colors vary from black to white, representing values ranging from 0 to 1. (c) and (i) show the results after the LR depth estimation step, and colors in the error maps vary from white to red, representing values ranging from 0 to 6 mm. (d),(e),(j),(k) show the results after the proposed uncertainty-aware GDSR step. (f) and (l) show the results after the uncertainty-aware refinement step.

with the conventional cascaded cost volume formulation on the DTU evaluation set using different input image resolutions. Specifically, the compared methods include CasMVSNet [15], GBi-Net [11], and MVSFormer [18], with results shown in Fig. 6. Recent methods have significantly improved both memory consumption and run-time compared to the naive cascaded MVS approach, CasMVSNet. However, MVSFormer, which integrates Vision Transformer for multi-view feature extraction, demonstrates a rapid increase in memory usage as input image resolution increases. On the other hand, GBi-Net reduces computational demands by restricting cost volume formulation to a small number of high-likelihood depth candidates. However, due to its iterative process, its inference time increases quickly as the resolution grows. In contrast, our method shows a slower increase in both memory consumption and run-time compared to other approaches. We attribute this efficiency to our super-resolution procedure, which operates in 2D space rather than the more computationally demanding 3D space. This design choice ensures a more scalable and efficient solution for HR depth estimation. Consequently, we conclude that the "cost volume + depth super-resolution" paradigm is advantageous for efficient HR depth estimation.

D. Ablation Study

To validate the contribution of each component in our method, we perform extensive ablation studies on the DTU dataset using identical settings to GBi-Net [11], namely,

TABLE V: Ablation Study of Each Component of our proposed SR-MVSNet on DTU evaluation dataset [7]

Model	CDE	DSR	Fusion	Depth (mm)		Point Cloud (mm)			Mem.(MB)	Run-Time (s)
				MAE	F.MAE	Acc.	Comp.	Overall		
Model-A	UCSNet	AHMF	Geo.	36.67	0.97	0.452	0.316	0.384	2784	0.34
Model-B	GBi-Net	AHMF	Geo.	21.12	0.82	0.372	0.352	0.362	**1769**	**0.33**
Model-C	GBi-Net	CFE+UMF(w/o CM)	Geo.	14.35	0.84	0.311	0.351	0.331	2508	0.37
Model-D	GBi-Net	CFE+UMF	Geo.	13.14	0.83	**0.306**	0.346	0.326	2508	0.37
Model-E	GBi-Net	CFE+UMF+Refine.	Photo.+Geo.	**10.37**	**0.55**	0.330	0.294	0.312	2839	0.46
Model-F	GBi-Net	CFE+UMF+Refine.	Dynamic.	**10.37**	0.62	0.386	0.228	0.307	2839	0.46
Model-G (Ours)	GBi-Net	CFE+UMF+Refine.	Iter. Fusion	**10.37**	0.57	0.387	**0.215**	**0.301**	2839	0.46

CDE, DSR, Fusion indicates the coarse LR depth estimation, depth super-resolution and multi-view depth fusion steps, respectively. Column Mem. (MB) measures the memory consumption of single HR depth estimation. CFE denotes the proposed Cost Feature Extraction module, UMF denotes the uncertainty-aware multi-modal feature fusion, CM denotes the confidence mask. Refine. denotes the proposed matching-based depth refinement module. Iter.F. denotes the proposed iterative depth fusion approach. The metric F.MAE stands for the filtered depth MAE.

1600×1152 image resolution and 5 input views. As for the photometric and geometric consistency filtering for final depth maps fusion, the photometric filtering threshold is set to $\xi_c = 0.8$, the pixel reprojection error threshold to $\xi_p = 1$, the depth reprojection error threshold to $\xi_d = 0.01$ and the minimum number of consistent views to $\xi_N = 3$. Results in Table V summarize a progressive series of seven model variants to isolate the impact of each component:

1 **Model-A**: Employs UCSNet [14] with default model settings for LR depth estimation, employs vanilla AHMF [27] for depth super-resolution, and employs default geometric consistency filtering (Geo. [3]) for multi-view depth maps fusion to final reconstruction.

2 **Model-B**: Replaces UCSNet [14] with GBi-Net [11].

3 **Model-C**: Incorporates the proposed Cost Feature Extraction (CFE, described in Subsection III-B1) and Uncertainty-aware Multi-modal Feature Fusion (UMF, described in Subsection III-B2) to vanilla AHMF, without confidence mask constraint.

4 **Model-D**: Further incorporates the Confidence Mask (CM, described in Subsection III-B3) constraint.

5 **Model-E**: Adds the proposed matching-based depth refinement module (Refine., described in Subsection III-C) and applies photometric and geometric consistency filterings (Photo.+ Geo. [3]) to get the final reconstruction.

6 **Model-F**: Replaces the photometric and geometric consistency filterings with dynamic consistency checking (Dynamic. [36]) to get the final reconstruction.

7 **Model-G (Our full model)**: Uses the proposed iterative filtering manner (described in Subsection III-E) to get the final reconstruction.

The performance of seven ablation models in terms of depth estimation, 3D reconstruction, memory usage, and average running time for single-depth-map inference are summarized in Table V. Initially, we simply connect GBi-Net [11] or UCSNet [14] for the LR depth estimation and classic GDSR network AHMF [27] for predicting full-resolution depth maps (Model-A and Model-B). We observe that the memory requirement and inference time are significantly lower than those of most MVS methods. However, the depth estimation and reconstruction performance are relatively poor. This can be attributed to the fact that AHMF primarily focuses on improving spatial resolution, assuming that LR measurements are clean and regularly sampled. However, LR depth estimation obtained from the cost volumes inevitably contains noise and outliers. Hence, directly connecting these two techniques makes it difficult to achieve good results. Comparing the performance of Model-A and Model-B, GBi-Net shows better performance. Furthermore, GBi-Net constructs cost volumes iteratively with only four depth candidates each time, making it both lightweight and efficient. Therefore, we select GBi-Net for the LR depth estimation in the subsequent experiments.

1) Benefit of Uncertainty-aware GDSR: As shown in Table V, the comparison between Model-B and Model-C demonstrates improved depth estimation and 3D reconstruction performance by incorporating the Cost Feature Extraction (CFE) and Uncertainty-aware Multi-modal Feature Fusion (UMF) modules in Model-C. This result confirms that local optimal LR outliers can be effectively managed during the depth super-resolution phase. Furthermore, Model-D integrates the Confidence Mask (CM) into the loss function, encouraging larger corrections in uncertain regions while minimizing residuals in multi-view consistent and reliable areas. As a result, Model-D exhibits a further performance enhancement over Model-C. When comparing Model-B and Model-D, we validated that the three modifications introduced in the uncertainty-aware GDSR module, including CFE, UMF, and CM, collectively improved depth estimation accuracy by 37.8% (MAE) and enhanced reconstruction performance by 9.9% (Overall), while requiring only an additional 739 MB of memory and increasing inference time by just 0.04 seconds.

Fig. 7 (i) and Fig. 7 (k) present the error comparisons for the traditional bicubic interpolation and our super-resolved depth. The results demonstrate that the uncertainty-aware GDSR module not only significantly reduces errors compared to bicubic interpolation but also requires less memory than most cascaded MVS methods.

We further visualize the extracted cost features, the input and output features of the original MMAF, and the proposed UMF in Fig. 8. It is observed that: 1) unimportant texture details are present in the extracted guidance feature; and 2) for LR depth input that contains errors, the original MMAF exhibits error-copying artifacts, whereas our UMF corrects these biased estimates based on extracted cost features. This enhancement is attributed to the proposed CFE and UMF, specifically tailored for matching-based LR depths.

2) Benefit of Matching-based Refinement: Models-A, B, C, and D solely produce HR depth maps without associated confidence maps, they rely only on geometric consistency

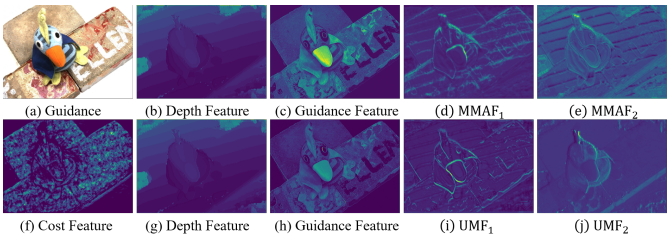


Fig. 8: Visualization of feature maps generated in the uncertainty-aware GDSR step. (b) and (g) are depth features derived from the LR depth map. (c) and (h) depict guidance features obtained from the HR guidance image. (f) illustrates the cost features generated by the proposed CFE, from the LR cost and probability volumes. (d) and (e) display feature maps fused by the original MMAF [27], without considering the uncertainty of LR depths. (i) and (j) showcase feature maps fused by the proposed UMF.

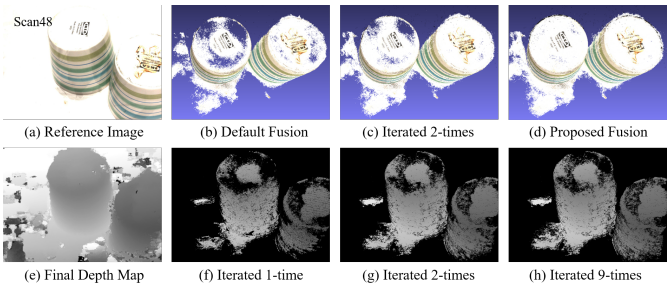


Fig. 9: Visualization of iteratively fused point clouds and filtered depth maps.

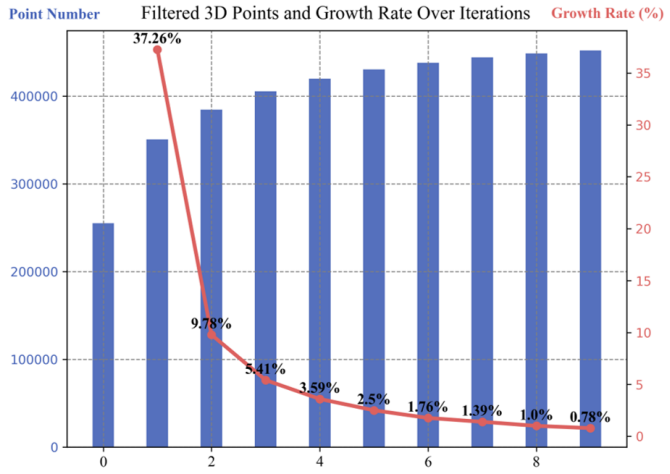


Fig. 10: The filtered depth point number and growth rate in each iteration of Fig. 9, with resolution 1600×1152.

filtering for multi-view depth fusion. In contrast, Model-E incorporates the proposed uncertainty-aware refinement module, yielding both refined HR depth maps and corresponding confidence maps. Therefore, conventional one-time photometric filtering based on confidence maps and geometric consistency filtering is employed for multi-view depth fusion. The lower depth MAE of Model-E in Table V indicates that our matching-based depth self-integration enhances the depth estimation accuracy. Meanwhile, Model-E applies strict outlier filtering based on predicted confidences, contributing to the lowest F.MAE and better overall quality of the reconstructed point clouds. Fig. 7 (k) and Fig. 7 (l) further provide the visual quality comparisons of the super-resolution and the refined results. It can be seen that the depth errors are further reduced in both rich and weak textured regions via our refinement. Finally, after incorporating the matching-based depth refinement step, SR-MVSNet requires 2839 MB of memory and 0.46 seconds for single-depth-map inference. The refinement step further enhances depth estimation accuracy by 21.1% (MAE) and improves reconstruction performance by 4.3% (Overall). With default depth fusion post-processing [3], Model-E achieves an overall score of 0.312 on the DTU evaluation set, comparable to most state-of-the-art MVS methods.

3) Benefit of Iterative Depth Fusion for 3D Reconstruction: As indicated in Table V, Models-E, F, and G maintain fixed LR depth estimation and guided depth super-resolution steps, enabling a fair comparison between the proposed iterative depth fusion, the default photometric and geometric consistency filtering [3] and the dynamic checking strategy [36] post-processing strategies. The results demonstrate that the proposed iterative depth fusion strategy achieves the best reconstruction completeness and overall quality. The results verify that our iterative depth fusion is effective and beneficial to achieving complete and accurate 3D reconstruction.

To illustrate the iterative fusion process, we present the iteratively fused point cloud (first row) and an iteratively filtered depth map (second row) in Fig. 9. In Fig. 9 (b) and Fig. 9 (f), where geometric consistency filtering is applied once, small and large holes appear in the point cloud and filtered depth map. Subsequently, in Fig. 9 (c) and Fig. 9 (g), depth self-integration is applied once, and geometric consistency filtering is applied twice, resulting in the filling of many small holes, although some larger textureless regions still exhibit unfilled portions in the depth map and fused point cloud. Finally, in Fig. 9 (d) and Fig. 9 (h), the proposed iterative depth self-integration and filtering are applied until converged, lead to a nearly complete fused point cloud. Simultaneously, the filtered depth errors (F.MAE) remain very small, as the geometric consistency checking conditions remain unchanged. The growth rate and filtered depth numbers after each iteration of Fig. 9, are visualized in Fig. 10. For this reference view, the valid depth points converge in nine iterations.

V. Conclusion

In this work, we presented SR-MVSNet, a novel deep MVS approach that achieves full-resolution depth estimation and 3D reconstruction by learning RGB-D structural consistency through the depth super-resolution technique. Unlike mainstream cascaded architectures that rely on coarse-to-fine depth sampling, struggling to balance performance with memory usage and varied depth ranges, we designed an uncertainty-aware guided depth super-resolution and a matching-based refinement module to avoid constructing high-resolution cost volumes. By leveraging mutual structural cues from RGB-D images and feature-matching metrics, SR-MVSNet delivers high-quality, full-resolution depth estimation while conserving computational resources. Finally, to generate the final point cloud, multi-view depth maps are fused using our proposed

iterative fusion post-processing strategy. Experiments on the DTU, Tanks & Temples, and ETH3D benchmarks demonstrate that our approach achieves competitive performance, particularly excelling on the ETH3D test set, which features wide-baseline, large-scale scenes with diverse depth ranges for each reference view. While not the highest across all metrics, our method offers an effective balance between reconstruction quality and computational efficiency, making it a practical solution for real-world reconstruction tasks. Ablation studies further confirm the effectiveness of our proposed contributions.

REFERENCES

[1] J. L. Schönberger, E. Zheng, J.-M. Frahm, and M. Pollefeys, "Pixelwise view selection for unstructured multi-view stereo," *IEEE/European Conference on Computer Vision*, p. 501–518, 2016.

[2] S. Fuhrmann, F. Langguth, and M. Goesele, "Mve-a multi-view reconstruction environment.," *GCH*, vol. 3, p. 4, 2014.

[3] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "MVSNet: Depth inference for unstructured multi-view stereo," *IEEE/European Conference on Computer Vision*, 2018.

[4] Y. Yao, Z. Luo, S. Li, T. Shen, T. Fang, and L. Quan, "Recurrent mvsnet for high-resolution multi-view stereo depth inference," *IEEE/Conference on Computer Vision and Pattern Recognition*, p. 5525–5534, 2019.

[5] C. Li, L. Zhou, H. Jiang, Z. Zhang, X. Xiang, H. Sun, Q. Luan, H. Bao, and G. Zhang, "Hybrid-MVS: Robust Multi-View Reconstruction with Hybrid Optimization of Visual and Depth Cues," *IEEE Transactions on Circuits and Systems for Video Technology*, p. 1–1, 2023.

[6] H. Xu, W. Chen, B. Sun, X. Xie, and W. Kang, "RobustMVS: Single Domain Generalized Deep Multi-view Stereo," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[7] 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, no. 2, p. 153–168, 2016.

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

[9] T. Schöps, J. L. Schönberger, S. Galliani, T. Sattler, and A. Geiger, "A Multi view Stereo Benchmark with High Resolution Images and Multi camera Videos," *IEEE/Conference on Computer Vision and Pattern Recognition*, 2017.

[10] 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," *IEEE/Conference on Computer Vision and Pattern Recognition*, 2020.

[11] Z. Mi, D. Chang, and D. Xu, "Generalized binary search network for highly-efficient multi-view stereo," *IEEE/Conference on Computer Vision and Pattern Recognition*, 2022.

[12] F. Wang, S. Galliani, C. Vogel, P. Speciale, and M. Pollefeys, "PatchmatchNet: Learned multi-view patchmatch stereo," *IEEE/Conference on Computer Vision and Pattern Recognition*, 2020.

[13] P. He, Y. Wang, Y. Wen, Y. Hu, and W. He, "ICV-Net: An identity cost volume network for multi-view stereo depth inference," *Pattern Recognition*, p. 111456, 2025.

[14] 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," *IEEE/Conference on Computer Vision and Pattern Recognition*, 2019.

[15] X. Gu, Z. Fan, S. Zhu, Z. Dai, and P. Tan, "Cascade Cost Volume for High-Resolution Multi-View Stereo and Stereo Matching," *IEEE/Conference on Computer Vision and Pattern Recognition*, 2020.

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

[17] K. T. Giang, S. Song, and S. Jo, "Curvature-guided dynamic scale networks for multi-view stereo," *International Conference on Learning Representations*, 2021.

[18] C. Cao, X. Ren, and Y. Fu, "MVSFormer: Multi-View Stereo by Learning Robust Image Features and Temperature-based Depth," *Transactions of Machine Learning Research*, 2023.

[19] 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, p. 8564–8576, 2022.

[20] W. Su, Q. Xu, and W. Tao, "Uncertainty guided multi-view stereo network for depth estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 11, p. 7796–7808, 2022.

[21] 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," *IEEE/International Conference on Computer Vision*, p. 17661–17670, 2023.

[22] R. Weilharter and F. Fraundorfer, "HAMMER: Learning Entropy Maps To Create Accurate 3D Models in Multi-View Stereo," *IEEE/Winter Conference on Applications of Computer Vision*, p. 3466–3475, 2024.

[23] K. Luo, T. Guan, L. Ju, H. Huang, and Y. Luo, "P-mvsnet: Learning patch-wise matching confidence aggregation for multi-view stereo," *IEEE/International Conference on Computer Vision*, p. 10452–10461, 2019.

[24] Y. Xue, J. Chen, W. Wan, Y. Huang, C. Yu, T. Li, and J. Bao, "Mvscrf: Learning multi-view stereo with conditional random fields," *IEEE/International Conference on Computer Vision*, p. 4312–4321, 2019.

[25] Q. Yan, Q. Wang, K. Zhao, B. Li, X. Chu, and F. Deng, "Rethinking disparity: a depth range free multi-view stereo based on disparity," *AAAI Conference on Artificial Intelligence*, vol. 37, no. 3, p. 3091–3099, 2023.

[26] Z. Zhong, X. Liu, J. Jiang, D. Zhao, and X. Ji, "Guided depth map super-resolution: A survey," *ACM Computing Surveys*, 2023.

[27] Z. Zhong, X. Liu, J. Jiang, D. Zhao, Z. Chen, and X. Ji, "High-resolution depth maps imaging via attention-based hierarchical multi-modal fusion," *IEEE Transactions on Image Processing*, vol. 31, p. 648–663, 2021.

[28] Y. Yang, Q. Cao, J. Zhang, and D. Tao, "CODON: on orchestrating cross-domain attentions for depth super-resolution," *International Journal of Computer Vision*, vol. 130, no. 2, p. 267–284, 2022.

[29] J. Yang, J. M. Alvarez, and M. Liu, "Non-parametric depth distribution modelling based depth inference for multi-view stereo," *IEEE/Conference on Computer Vision and Pattern Recognition*, 2022.

[30] J. Zhu, B. Peng, W. Li, H. Shen, Q. Huang, and J. Lei, "Modeling long-range dependencies and epipolar geometry for multi-view stereo," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 6, p. 1–17, 2023.

[31] 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," *IEEE/European Conference on Computer Vision*, p. 573–591, 2022.

[32] J. Wu, R. Li, H. Xu, W. Zhao, Y. Zhu, J. Sun, and Y. Zhang, "GoMVS: Geometrically consistent cost aggregation for multi-view stereo," *IEEE/Conference on Computer Vision and Pattern Recognition*, 2024.

[33] H. Li, Y. Guo, X. Zheng, and H. Xiong, "Learning deformable hypothesis sampling for accurate patchmatch multi-view stereo," *AAAI Conference on Artificial Intelligence*, vol. 38, no. 4, p. 3082–3090, 2024.

[34] J. Li, Z. Lu, Y. Wang, J. Xiao, and Y. Wang, "NR-MVSNet: Learning multi-view stereo based on normal consistency and depth refinement," *IEEE Transactions on Image Processing*, 2023.

[35] S. Zhang, W. Xu, Z. Wei, L. Zhang, Y. Wang, and J. Liu, "ARAI-MVSNet: A multi-view stereo depth estimation network with adaptive depth range and depth interval," *Pattern Recognition*, vol. 144, p. 109885, 2023.

[36] 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," *IEEE/International Conference on Computer Vision*, 2020.

[37] K.-H. Lo, Y.-C. F. Wang, and K.-L. Hua, "Edge-preserving depth map upsampling by joint trilateral filter," *IEEE Transactions on Cybernetics*, vol. 48, no. 1, p. 371–384, 2018.

[38] Z. Shi, Y. Chen, E. Gavves, P. Mettes, and C. G. M. Snoek, "Unsharp mask guided filtering," *IEEE Transactions on Image Processing*, p. 7472–7485, 2021.

[39] J. Yang, X. Ye, K. Li, C. Hou, and Y. Wang, "Color-guided depth recovery from RGB-D data using an adaptive autoregressive model," *IEEE Transactions on Image Processing*, p. 3443–3458, 2014.

[40] D. Ferstl, C. Reinbacher, R. Ranftl, M. Ruether, and H. Bischof, "Image guided depth upsampling using anisotropic total generalized variation," *IEEE/International Conference on Computer Vision*, 2013.

[41] R. D. Lutio, A. Becker, S. D'Aronco, S. Russo, J. D. Wegner, and K. Schindler, "Learning graph regularisation for guided super-resolution," *IEEE/Conference on Computer Vision and Pattern Recognition*, p. 1979–1988, 2022.

[42] I. Marivani, E. Tsiligianni, B. Cornelis, and N. Deligiannis, "Multimodal deep unfolding for guided image super-resolution," *IEEE Transactions on Image Processing*, p. 8443–8456, 2020.

[43] F. Gao, X. Deng, M. Xu, J. Xu, and P. L. Dragotti, "Multi-modal convolutional dictionary learning," *IEEE Transactions on Image Processing*, p. 1325–1339, 2022.

[44] Y. Zuo, Q. Wu, Y. Fang, P. An, L. Huang, and Z. Chen, "Multi-scale frequency reconstruction for guided depth map super-resolution via deep residual network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 2, p. 297–306, 2019.

[45] B. Kim, J. Ponce, and B. Ham, "Deformable kernel networks for joint image filtering," *International Journal of Computer Vision*, p. 579–600, 2021.

[46] X. Song, Y. Dai, D. Zhou, L. Liu, W. Li, H. Li, and R. Yang, "Channel attention based iterative residual learning for depth map super-resolution," *IEEE/Conference on Computer Vision and Pattern Recognition*, p. 5631–5640, 2020.

[47] 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, p. 199–214, 2023.

[48] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. Huang, "Free-form image inpainting with gated convolution," *IEEE/International Conference on Computer Vision*, 2019.

[49] S. Song, K. G. Truong, D. Kim, and S. Jo, "Prior depth-based multi-view stereo network for online 3D model reconstruction," *Pattern Recognition*, vol. 136, p. 109198, 2023.

[50] S. Galliani, K. Lasinger, and K. Schindler, "Massively parallel multiview stereopsis by surface normal diffusion," *IEEE/International Conference on Computer Vision*, 2015.

[51] H. Yi, Z. Wei, M. Ding, R. Zhang, Y. Chen, G. Wang, and Y.-W. Tai, "Pyramid multi-view stereo net with self-adaptive view aggregation," *IEEE/European Conference on Computer Vision*, p. 766–782, 2020.

[52] R. Zhao, X. Han, X. Guo, L. Kuang, X. Yang, and F. Sun, "Exploring the point feature relation on point cloud for multi-view stereo," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 11, p. 6747–6763, 2023.

[53] TnT online reconstruction results: https://www.tanksandtemples.org/details/7806/, https://www.tanksandtemples.org/details/8866/, https://www.tanksandtemples.org/details/9228/, https://www.tanksandtemples.org/details/9216/. Evaluation service available at https://www.tanksandtemples.org/leaderboard/ provided by A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun, "Tanks and temples: Benchmarking large-scale scene reconstruction," 2017.

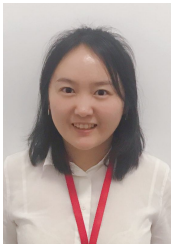
[54] ETH3D online reconstruction results: https://www.eth3d.net/result_details?id=1346, https://www.eth3d.net/result_details?id=1807. Evaluation service available at https://www.eth3d.net/high_res_multi_view/ provided by Schps, T and Schnberger, J. L and Galliani, S. and Sattler, T. and Geiger, A., "A multi view stereo benchmark with high resolution images and multi camera videos," 2017.

[55] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," *IEEE/International Conference on Computer Vision*, 2015.

[56] E. Ilg, O. Cicek, S. Galesso, A. Klein, O. Makansi, F. Hutter, and T. Brox, "Uncertainty estimates and multi-hypotheses networks for optical flow," *IEEE/European Conference on Computer Vision*, p. 652–667, 2018.

[57] P. Schröppel, J. Bechtold, A. Amiranashvili, and T. Brox, "A benchmark and a baseline for robust multi-view depth estimation," *International Conference on 3D Vision*, p. 637–645, 2022.

VI. Biography Section



Yimei Liu received the B.Sc. degree from Xidian University, Xian, China, in 2013 and M.E. degree from University Jean Monnet, Saint-Etienne, France, in 2015. She is currently pursuing the Ph.D. degree with Ocean University of China. Her research interests include computer vision and 3D reconstruction.

Jingchao Cao received the B.S. degree in electronic engineering from Xidian University, Xi'an, China, in 2014, and M.S. and Ph.D. degrees in computer science from the City University of Hong Kong, Hong Kong, in 2016 and 2022, respectively. He is currently a Lecturer with the Department of Information Science and Technology, Ocean University of China, Qingdao, China. His research interests include computer vision, machine learning, image quality assessment, and image enhancement.

Hao Fan received his B.Sc., M.E., and Ph.D. degrees from the Department of Computer Science and Technology at Ocean University of China, Qingdao, China, in 2012, 2014 and 2019, respectively. He is currently a Lecture with the Department of Information Science and Technology, Ocean University of China. His research interests include computer vision, 3D reconstruction and underwater image processing.

Junyu Dong received the B.Sc. and M.Sc. degrees from the Department of Applied Mathematics, Ocean University of China, in 1993 and 1999, respectively, and the Ph.D. degree in image processing from the Department of Computer Science, Heriot-Watt University, U.K., in November 2003. He joined the Ocean University of China in 2004, where he is currently a Professor and the Head of the Department of Information Science and Technology. His research interests include machine learning, big data, computer vision, and underwater image processing.

Sheng Chen (IEEE Life Fellow) received his BEng degree from the East China Petroleum Institute, Dongying, China, in 1982, and his PhD degree from the City University, London, in 1986, both in control engineering. In 2005, he was awarded the higher doctoral degree, Doctor of Sciences (DSc), from the University of Southampton, Southampton, UK. From 1986 to 1999, He held research and academic appointments at the Universities of Sheffield, Edinburgh and Portsmouth, all in UK. Since 1999, he has been with the School of Electronics and Computer Science, the University of Southampton, UK, where he holds the post of Professor in Intelligent Systems and Signal Processing. Dr Chen's research interests include machine learning, neural networks and wireless communications. He has published over 700 research papers. Professor Chen has 19,000+ Web of Science citations with h-index 61 and 38,000+ Google Scholar citations with h-index 83. Dr. Chen is a Fellow of the United Kingdom Royal Academy of Engineering, a Fellow of Asia-Pacific Artificial Intelligence Association and a Fellow of IET. He is one of the original ISI highly cited researcher in engineering (March 2004).