Geometry-Enhanced Attentive Multi-View Stereo for Challenging Matching Scenarios

Yimei Liu, Qin Cai, Congcong Wang, Jian Yang, Hao Fan, Junyu Dong, *Member, IEEE*,
Sheng Chen, *Fellow, IEEE* .

***Abstract*—Deep networks have made remarkable progress in Multi-View Stereo (MVS) task in recent years. However, the problem of finding accurate correspondences across different views under ill-posed matching situations remains unresolved and crucial. To address this issue, this paper proposes a Geometry-enhanced Attentive Multi-View Stereo (GA-MVS) network, which can access multi-view consistent feature representation and achieve accurate depth estimation in challenging situations. Specifically, we propose a geometry-enhanced feature extractor to explore illumination-invariant geometric features and incorporate them with common texture features to improve matching accuracy when dealing with view-dependent photometric effects, such as shadow and specularity. Then, we design a novel attentive learning framework to explore per-pixel adaptive supervision, effectively improving the depth estimation performance of textureless regions. The experimental results on the DTU and Tanks & Temples benchmarks demonstrate that our method achieves state-of-the-art results compared to other advanced MVS models.**

***Index Terms*—Multi-View Stereo, 3D Reconstruction, Depth Estimation, Geometric features, Deep Learning.**

I. Introduction

Multi-View Stereo (MVS) aims to densely reconstruct the 3D geometry of a scene by utilizing multiple-view images and corresponding camera parameters. MVS is an essential technique for 3D reconstruction and has been extensively studied for decades due to its wide range of applications, including augmented reality [1], scene reconstruction [2]–[4], photogrammetry [5], [6] and cartography [7]–[11]. A prevalent paradigm of MVS is to estimate the depth map for each input image from multiple observations and then merge these multi-view depth maps to generate dense 3D reconstructions. So far, numerous MVS methods have been proposed using this paradigm, ranging from early traditional methods [12]–[14] to recent deep learning-based methods [15]–[19].

The traditional MVS methods, such as OpenMVS [13] and COLMAP [14], generally rely on hand-crafted features and matching metrics to evaluate multi-view photo-consistency, enabling accurate depth estimation in well-textured, ideal Lambertian scenes [20], [21]. However, these methods encounter difficulties in regions with shadows, reflections, and lack of texture, where matching problems become challenging and ill-posed. These ill-posed matching issues can be categorized into two main aspects: the variation in appearance textures among multiple views due to view-dependent photometric effects, and ambiguous matching results caused by homogeneous textureless regions. Both aspects negatively impact the robustness of multi-view matching. To enhance performance, recent deep learning-based methods [15]–[17], [22]–[25] employ Convolutional Neural Networks (CNNs) to incorporate semantic information for more reliable matching, leading to improved performance on various MVS benchmarks [26]–[29]. Some approaches, such as the geometry-based methods [30]–[33] and the attention-based methods [17], [34]–[37], introduce stable geometric clues and finely designed attention mechanisms to alleviate ill-posed matching issues. These methods perform well in partially challenging situations, namely, either for view-dependent photometric effects only or for homogeneous textureless areas only. Furthermore, attention mechanisms introduce high computation complexity. More importantly, these approaches do not fully leverage valuable ground truth depths, as they only impose constraints on the final regressed or classified depths, indirectly affecting the early-stage extracted features. This ambiguity implicitly adds the difficulty to robust feature learning in the MVS task.

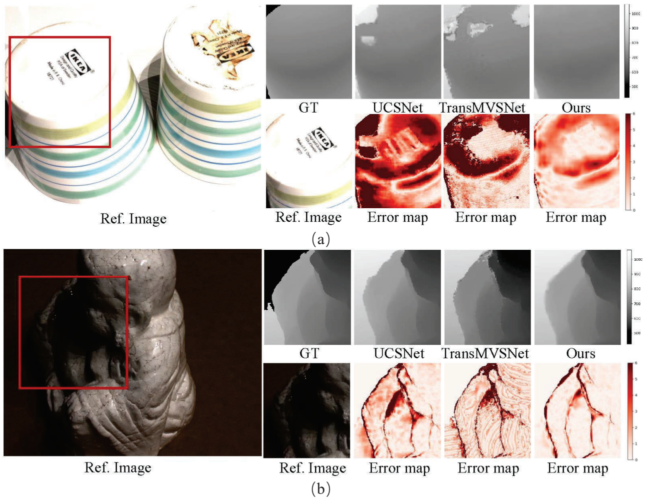
To tackle the aforementioned problems, we propose an advanced framework called Geometry-enhanced Attentive MVS (GA-MVS). GA-MVS comprises two crucial steps: a geometry-enhanced feature extractor and an attentive learning framework. The geometry-enhanced feature extractor aims to obtain 3D consistent feature representations, while the attentive learning framework uses the 3D consistent features to produce accurate depth estimations in regions with varying texture richness. Specifically, in the proposed geometry-enhanced feature extractor, a Geometry-Aware Module (GAM) is first introduced to extract illumination-invariant geometric feature from common multi-level texture feature. Importantly, in this step, we incorporate accurate depth variance constraints to guide the model's attention towards regions with varied depths, which contains valuable geometric information. Then, the Feature Fusion Module (FFM) is designed to effectively integrate the acquired geometric feature with initial multi-level texture feature, thus producing the final geometry-enhanced representations. In the proposed attentive learning framework, we first estimate a reference depth map and an aligned adaptive

This work was supported in part by National Science Foundation of China (Grant No. 42106193, 41927805). Our implementation is available at https://github.com/yimei910110/GA-MVS. (*Corresponding authors*: Junyu Dong, Hao Fan.)

Y. Liu, Q. Cai, J. Yang, 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, cq@ouc.edu.cn, yangjian@stu.ouc.edu.cn, fanhao@ouc.edu.cn, dongjunyu@ouc.edu.cn).

C. Wang is with the Department of Computer Science and Engineering, Tianjin University of Technology, Tianjin, China (email: congcong_wang@yeah.net)

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).



(a)

(b)

Fig. 1: Performance comparison of learning-based MVS methods on shadow and textureless regions: the baseline model [16], the state-of-the-art transformer-based approach TransMVSNet [34], and our proposed GA-MVS. By incorporating **geometry-enhanced features** and **attentive learning framework**, our method achieves more accurate depth estimates in challenging matching scenarios.

attention map using the baseline depth estimation network and a lightweight attention network. The latter provides weights for the pixel-wise attention-balanced loss, which is composed of the depth loss and gradient loss. Consequently, pixels in textureless regions are assigned higher gradient loss weights due to their higher matching difficulty. As illustrated in Fig. 1, conventional smooth L_1 loss may lead to biased estimates in these challenging areas. In contrast, our attentive learning framework learns from the attention-balanced loss, enabling smooth depth estimates in textureless regions and accurate estimates in general textured regions.

Our main contributions are summarized as follows:

- A geometry-enhanced feature extractor is proposed to explore 3D consistent features that are robust to view-dependent photometric effects. It is implemented by introducing reliable constraints to help the model explore real geometric clues with varied depths and incorporate them with discriminative texture features.
- An attentive learning framework is proposed to explore adaptive constraint via attention-balanced loss. This learning schema can be built upon various MVS networks and is crucial for enabling accurate depth estimations in regions with varying texture richness.
- We verify the effectiveness of the proposed geometry-enhanced features and attentive learning framework on two benchmarks: DTU and Tanks & Temples [26], [27]. The results demonstrate that our method can significantly enhance MVS reconstruction performance and outperforms existing state-of-the-art methods.

The rest of this paper is organized as follows. Section II provides an overview of related works. Section III presents a comprehensive explanation of our proposed methodology, which includes the geometry-enhanced feature extractor and the attentive learning framework. Section IV compares the reconstruction results of our proposed GA-MVS with those of state-of-the-art models on two MVS benchmarks and thoroughly analyzes the performance. Furthermore, Subsection IV-D validates the effectiveness of the two components of GA-MVS and presents more insights on their performance improvements through a series of ablation experiments. Finally, the paper concludes in Section V.

II. Related Work

According to the taxonomy provided in [38], MVS methods can be generally divided into surface evolution-based, voxel-based, point cloud-based, and depth-based methods. Since our method falls in the category of depth-based methods, this section will focus on the review of depth-based MVS methods, attention mechanisms and geometric clues that have been applied in the stereo vision field.

A. Learning-based MVS

In contrast to traditional MVS methods, learning-based MVS methods have made significant progress in recent years, owing to their robust feature representation and adaptive matching measurement enabled by CNNs. As a groundbreaking depth map-based method, MVSNet [15] proposed an end-to-end pipeline to support depth estimation based on 3D cost volumes and 3D-CNNs. However, the memory requirement grows explosively with the input image resolution and depth hypotheses augment. To reduce memory consumption, two types of strategies have been proposed. Recurrent approaches, such as R-MVSNet [39] and D2HC-RMVSNet [40], introduce gated recurrent units (GRUs) and long short-term memory networks (LSTMs) for cost volume regularization, which trade off time consumed with low memory costs. On the other hand, coarse-to-fine approaches, like CasMVSNet [41], UCSNet [16] and CVP-MVSNet [42], formulate cascaded cost volumes to reduce computational complexity. Additionally, other researchers have explored improving the pipeline performance through other aspects. PatchmatchNet [43] and GBINet [44] proposed more effective depth hypothesis generation strategies for cost volume construction. UGNet [44], Vis-MVSNet [44], U-MVS [45] and NP-CVP-MVS [46] achieved uncertainty-guided depth map estimation by exploring pixel-wise depth probability distribution modelings.

Enlightened by these works, our proposed GA-MVS constructs cascade cost volumes to estimate high-resolution depth maps from coarse to fine. In particular, to address two major issues in the MVS task, i.e., view-dependent photometric effects and matching ambiguity associated with textureless regions, we propose a geometric-enhanced feature extractor and an attentive learning framework, which can be combined with multiple coarse-to-fine MVS methods to boost their depth estimation performances.

B. Attention Mechanisms for Learning-based MVS

The attention mechanism has been widely investigated in various visual tasks. Some researchers have focused on developing plug-and-play attention modules for general feature representation [47]–[50]. Specifically, SENet [47] proposed a channel-wise attention mechanism that highlights critical channels for downstream tasks. The work of [48] proposed a spatial-wise attention mechanism for capturing long-range feature dependencies. CBAM [49] and BAM [50] proposed mixed attention mechanisms that consider both channel and spatial-wise feature interactions.

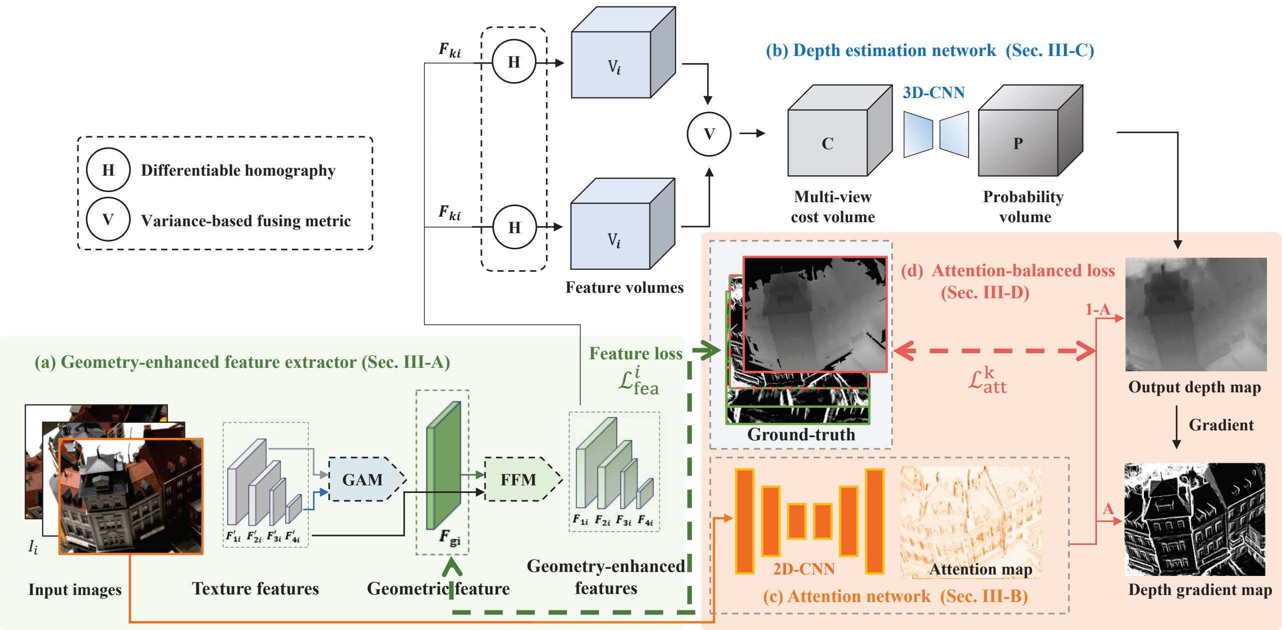
Recently, attention mechanisms have been explored in multiple MVS methods. MVSTR [37] and AACVP-MVSNet [35] leveraged attention mechanisms to learn more reliable features than conventional Feature Pyramid Networks (FPNs). AttMVS [17] and TransMVSNet [34] proposed attention-guided regularization modules instead of variance-based feature fusion metrics and 3D-CNNs regularization steps. Moreover, MVSTER [36] and RayMVSNet [18] proposed limiting attention associations within the epipolar line to reduce computation.

Unlike these previous works that applied attention mechanisms for cost volume construction or regularization, our proposed GA-MVS generates an attentive spatial attention map, which affects the weights of per-pixel attention-balanced loss and provides appropriate penalty strategies for different areas with varied matching difficulties, thereby improving the accuracy of depth estimates in challenging textureless regions.

C. Geometric Clues for Learning-based MVS

Although convolutional features are commonly employed to describe the points and construct matching costs in current MVS networks, they contain less geometric information since the learned kernels are directly impacted by appearance texture variance. As a view transforms, the appearance textures may change while the geometric characteristics tend to remain more stable [51]. Therefore, the works [30]–[32] focused on combining textured and geometric clues for accurate matching and cost aggregation. Specifically, LSP [30] introduced learning-based structure features for deep stereo-matching networks, providing complementary information to CNN-based texture features. EdgeStereo [31] incorporated an edge detection subnetwork to explore edge clues and serve them as important guidances for disparity learning. CDS-MVSNet [32] proposed to calculate pixel-wise normal curvatures along the epipolar line, which can be used to access reliable features for robust multi-view matching. All these existing works learn geometric clues of the scene without direct and reliable constraints. Therefore, the extracted geometric clues are susceptible to realistic illumination and view angle effects, such as pseudo textures caused by shadow or specularity. These pseudo textures cause inconsistent feature representation and erroneous matching results, thereby detrimental to accurate matching.

Motivated by these existing researches and aiming to address their weakness, we propose a novel GAM to extract a one-channel geometric feature. In particular, we leverage reliable depth constraints to facilitate the model to learn real geometric clues of the scene. The obtained geometric feature, together with conventional multiscale texture feature, are effectively integrated to obtain the final robust representation, which is beneficial for mitigating the inevitably negative view-dependent photometric effects on the accuracy of matching results.



F_{ki}
H
V_i
F_{ki}
H
V_i
Feature volumes
H Differentiable homography
V Variance-based fusing metric
V
(b) Depth estimation network (Sec. III-C)
3D-CNN
C
P
Multi-view cost volume
Probability volume
(d) Attention-balanced loss (Sec. III-D)
1-A
$\mathcal{L}_{att}^{k}$
Output depth map
Ground-truth
Gradient
A
Depth gradient map
(a) Geometry-enhanced feature extractor (Sec. III-A)
Feature loss
$\mathcal{L}_{fea}^{i}$
I_i
F'_{1i} F'_{2i} F'_{3i} F'_{4i}
GAM
F_{gi}
FFM
F_{1i} F_{2i} F_{3i} F_{4i}
Input images
Texture features
Geometric feature
Geometry-enhanced features
2D-CNN
Attention map
(c) Attention network (Sec. III-B)

Fig. 2: The overall architecture of GA-MVS. The main components include: (a) Geometry-enhanced feature extractor, which outputs feature maps of input images hierarchically at multiple resolutions, thus realizing depth estimation in a coarse-to-fine manner. (b) Depth estimation network, which conducts on each stage based on the cascaded cost volumes pipeline. (c) Attention network, which generates adaptive multi-scale attention maps aligned with estimated depth maps and provides weights for (d) per-pixel attention-balanced loss. The subindices $i \in \{0, 1, \cdots, N\}$ denotes the input images, while subindices $k \in \{1, 2, 3, 4\}$ denotes the levels of features.

III. Methodology

Fig. 2 depicts the overall architecture of the proposed GA-MVS. We first introduce the proposed geometry-enhanced feature extractor in Subsection III-A. Then we present our attention network in Subsection III-B and our depth estimation network in Subsection III-C. Next we define our loss functions in Subsection III-D, which include the novel attention-balanced loss and feature loss. To investigate the effectiveness of our method, we adopt the cascade MVS method UCSNet [16] as the baseline to predict final depth maps.

A. Geometry-Enhanced Feature Extractor

Given the reference image I_0 and its neighboring source images $\{I_i\}_{i=1}^{N}$, before estimating the reference depth map, we encode the input images into multi-scale feature maps, as illustrated in the part (a) of Fig. 2. Specifically, we first employ the common FPN [52] to extract the initial multi-scale texture features F'_{ki} ($k \in \{1,2,3,4\}, i \in \{0,1,\cdots,N\}$), where subindices k denote stages, corresponding to spatial resolutions of $W/2^{k-1} \times H/2^{k-1}$. Here, W and H are the width and height of input images. To access illumination-invariant geometric features, we introduce the GAM in Subsection III-A1, which explores stable one-channel geometric features F_{gi} ($i \in \{0,1,\cdots,N\}$) from the multi-scale texture features with aligned depth gradient maps as constraints. Then, we present the FFM in Subsection III-A2, which integrates the geometric feature from GAM with the initial texture features at each level to obtain the final geometry-enhanced representation F_{ki}.

1) Geometry-Aware Module GAM: Low-level features contain rich geometric clues but also introduce irrelevant texture details. High-level semantic information is needed to facilitate the exploration of real geometric clues. Therefore, we propose to incorporate low-level features F'_{1i} and high-level features F'_{4i} to model the real one-channel geometric feature. We further enforce the output one-channel geometric feature by constraining it with the calculated ground-truth depth gradient map, a kind of accurate 3D-geometric clue. The geometric features are trained implicitly by backpropagation when training the end-to-end networks with the proposed feature loss defined in Subsection III-D1.

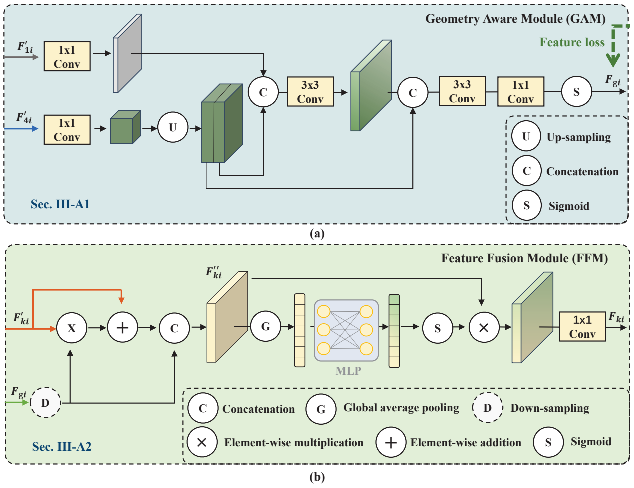


Fig. 3: (a) Illustration of GAM, which explores illumination-invariant geometric feature F_{gi}, where the green dashed arrow indicates the proposed feature loss. (b) Illustration of FFM, which integrates geometric feature F_{gi} into initial texture features F'_{ki} to get stable and discriminative representation F_{ki}. Subindices $i \in \{0,1,\cdots,N\}$ denote the input images, while subindices $k \in \{1,2,3,4\}$ denote the levels of features.

The procedure of this GAM is illustrated in Fig. 3 (a). Specifically, we first apply 1×1 convolution layers to change the channels of F'_{1i} and F'_{4i}. Then, the feature F'_{1i} and the up-sampled F'_{4i} are concatenated two times for integration. Finally, the one-channel geometric feature is obtained through two convolution layers and a sigmoid function. This GAM can be formulated as follows:

$$F_{gi} = f_{gam}\big((F'_{1i}, F'_{4i}); \theta_{gam}\big), \tag{1}$$

where $f_{gam}(\cdot; \theta_{gam})$ denotes the mapping of the GAM with the learnable parameters θ_{gam}. GAM is a simple yet effective module to extract geometric features. The feature loss used to train the GAM helps the network focusing on the region with varied depths, which contains real geometric clues. This will become clear in Subsection III-D1

2) Feature Fusion Module FFM: Noting different feature channels focusing on different image regions, we first modulate the initial texture features using the obtained geometric feature in the channel dimension. Then, using the channel attention mechanism [47], we explore the cross-channel interaction and further enhance critical ones for matching.

The procedure of our FFM is illustrated in Fig. 3 (b). Its input features include F'_{ki} and F_{gi}, which are multi-level texture features from the vanilla FPN and geometric features from the GAM. We first perform the element-wise multiplication and element-wise addition skip connection between them. The channels containing varied depths can be enhanced while the others remain unchanged. Then, we concatenate the processed feature with the geometric feature to obtain the initially fused feature F''_{ki}. This process can be formulated as follows:

$$F''_{ki} = C\big((F'_{ki} \otimes D(F_{gi})) \oplus F'_{ki}, D(F_{gi})\big), \tag{2}$$

where $D(\cdot)$ denotes the down-sampling operation, applied to adjust the geometric feature to the corresponding resolution, $C(\cdot)$ represents the concatenation operation, $\otimes$ indicates element-wise multiplication, and $\oplus$ indicates element-wise addition. Then, we enhance the critical feature channels by performing the effective channel attention mechanism introduced in [47]. Finally, we adjust the channel number through 1×1 convolutions to get geometry-enhanced features F_{ki}. The process can be formulated as follows:

$$F_{ki} = f_{ffm}\Big(\sigma\big(MLP(G(F''_{ki}))\big) \otimes F''_{ki}; \theta_{ffm}\Big), \tag{3}$$

where $\sigma(\cdot)$ denotes the sigmoid function, $G(\cdot)$ denotes the channel-wise global average pooling operation, $MLP(\cdot)$ indicates a two-layer Multi-Layer-Perceptron (MLP) and $f_{ffm}(\cdot; \theta_{ffm})$ represents 1×1 convolution layers with the learnable parameters θ_{ffm}. The applied channel attention mechanism can highlight critical channels and suppress redundant ones, thereby enhancing robust feature representation. As will be shown later in Fig. 11 of Subsection IV-D, in comparison to the initial texture features F'_{ki}, our geometry-enhanced features F_{ki} are able to effectively attenuate view-dependent photometric effects, thus benefiting the robustness of the multi-view matching.

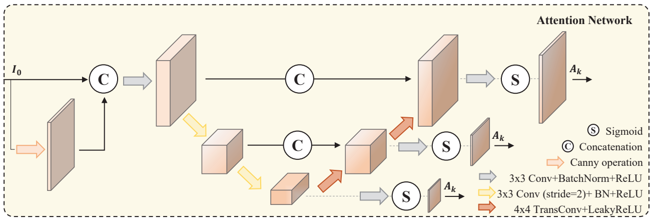


Fig. 4: Illustration of the attention network, which outputs multi-scale attention maps $\{A_k\}_{k=1}^{4}$ from the reference image I_0, for per-pixel attention-balanced loss calculation. The resolution of A_k is $1 \times \frac{W}{2^k} \times \frac{H}{2^k}$, with input image of size $W \times H$.

B. Attention Network

To help the model learn the pattern of depth estimation in textureless regions, we propose an attentive learning framework that provides adaptive weights for the attention-balanced loss calculation, detailed in Subsection III-D2. Fig. 4 illustrates the proposed attention network. Since the edges are generally associated with textureless, oppositely, the attention network first contains an edge-detection layer, calculated by applying the canny operator and three-layer convolutions on input I_0. Then, the concatenated I_0 and its edge $f_{edge}(I_0)$ pass through a lightweight UNet structured 2D CNN network, consisting of three-layer convolutions, three-layer deconvolutions, and sigmoid non-linearities, denoted as f_{att}. Finally, we obtain the adaptive multi-scale attention maps $\{A_k\}_{k=1}^{4}$ from the reference image I_0 formulated as follows:

$$A_k = f_{att}\big(I_0, f_{edge}(I_0); \theta_{att}\big), \tag{4}$$

where $A_k \in \mathbb{R}^{1 \times H/2^{k-1} \times W/2^{k-1}}$, with the same resolution as the multi-scale depth estimates. The learnable parameters in the attention network are denoted as θ_{att}.

C. Depth Estimation Network

We adopt a typical depth estimation network of [8], [16], [41], [53], After the feature extraction step, the multi-stage depth maps aligned to the reference view are estimated from coarse to fine. For the k-th stage depth estimation, subindex denoting the stage is omitted for simplicity. Based on a set of depth hypotheses $\{d_j\}_{j=1}^{D}$ and the pre-calculated camera intrinsic and extrinsic matrices $\{K_i, T_i\}_{i=0}^{N}$, we construct a set of feature volumes $\{V_i\}_{i=1}^{N}$ by differentiable warping of the geometry-enhanced source features $\{F_i\}_{i=1}^{N}$ into the reference view as follows:

$$V_i = f_p\big(F_i, p(d_j)\big) = f_p\big(F_i, d_j K_i T_i T_0^{-1} K_0^{-1}\big), \tag{5}$$

where $f_p(\cdot,\cdot)$ denotes the differentiable bilinear interpolation of source feature map at the normalized pixel coordinates $p(d_j)$, and $p(d_j)$ are calculated by homography between the i-th source and the reference image at a depth set $\{d_j\}_{j=1}^{D}$. Then, multiple feature volumes are aggregated to one 3D cost volume C by the variance-based fusing metric, which is calculated as:

$$C = \frac{1}{N}\sum_{i=0}^{N}\big(V_i - \overline{V}\big)^2, \tag{6}$$

where $\overline{V}$ denotes the average feature volume. The essence is using the variance of warped source features to measure the confidence of a depth hypothesis. For the depth hypothesis with high confidence, the variance of warped source features should be small, because they represent the same 3D point in space, and vice versa. Next, the 3D cost volume C is regularized by a 3D-CNN and transformed into a probability volume $P \in R^{D \times H \times W}$ given by:

$$P = f_{dr}\big(C; \theta_{dr}\big), \tag{7}$$

where $f_{dr}\big(\cdot; \theta_{dr}\big)$ denotes the mapping of the 3D-CNN with the learnable parameters θ_{dr}. Finally, the depth map $\widetilde{\mathbf{D}}$ at the current stage is regressed via soft-argmax operation, with the depth value at each pixel $\mathbf{p}$ computed as follows:

$$\widetilde{\mathbf{d}}_{\mathbf{p}} = \sum_{j=1}^{D} d_j P_j(\mathbf{p}), \tag{8}$$

with the probability volume $P_j \in R^{1 \times H \times W}$. The estimated depth map $\widetilde{\mathbf{d}}_{\mathbf{p}}$ is up-sampled to fit the spatial resolution. Then, a set of depth hypotheses are generated uniformly in the variance-based confidence interval [16], centering on the recent outcome, for higher resolution depth map estimation.

D. Loss Function

The overall loss comprises two components: the feature loss of multiple views and the attention-balanced loss of multiple stage depth estimates, formulated as follows:

$$\mathcal{L}_{total} = \alpha \sum_{i=0}^{N} \mathcal{L}_{fea.}^{i} + \sum_{k=1}^{4} \mathcal{L}_{att.}^{k}. \tag{9}$$

The hyper-parameter α controls the relative importance of the two components, which is set to be 0.5 in our experiments.

1) Feature loss: The feature loss $\mathcal{L}_{fea.}^{i}$ is applied to the geometric feature maps from the GAM and the corresponding ground-truth depth gradient maps. Both maps are pre-scaled into the range $[0,\ 1]$. Then, the mean squared error function is utilized to calculate the loss.

2) Attention-balanced loss: The stage index k is omitted for notational simplicity. The network parameters θ are optimized by minimizing the attention-balanced loss of pixels in the ground-truth reference depth valid region Ω:

$$\mathcal{L}_{att.} = \sum_{\mathbf{p}\in\Omega} \omega_{\mathbf{p}} \mathcal{L}_g\big(\widetilde{\mathbf{d}}_{\mathbf{p}}, \mathbf{d}_{\mathbf{p}}^{gt}\big) + \lambda(1-\omega_{\mathbf{p}})\mathcal{L}_d\big(\widetilde{\mathbf{d}}_{\mathbf{p}}, \mathbf{d}_{\mathbf{p}}^{gt}\big), \tag{10}$$

where $\omega_{\mathbf{p}}$ represents the predicted attention value at pixel $\mathbf{p}$, taken from the output attention map A_k at stage k, and λ is a protective threshold to prevent insufficient penalty on the output depth map, which is set to 8 in our experiments.

The first part of the attention-balanced loss, $\mathcal{L}_g\big(\widetilde{\mathbf{d}}_{\mathbf{p}}, \mathbf{d}_{\mathbf{p}}^{gt}\big)$, describes the gradient loss between the estimated depth $\widetilde{\mathbf{d}}_{\mathbf{p}}$ and the ground-truth depth $\mathbf{d}_{\mathbf{p}}^{gt}$ at pixel $\mathbf{p}$, given by

$$\mathcal{L}_g\big(\widetilde{\mathbf{d}}_{\mathbf{p}}, \mathbf{d}_{\mathbf{p}}^{gt}\big) = \Big\| g\big(\widetilde{\mathbf{d}}_{\mathbf{p}(x,y)}, \varepsilon\big), g\big(\mathbf{d}_{\mathbf{p}(x,y)}^{gt}, \varepsilon\big) \Big\|_2, \tag{11}$$

where (x, y) are the pixel coordinates of $\mathbf{p}$. We define the depth gradient $g\big(\mathbf{d}_{\mathbf{p}(x,y)}, \varepsilon\big)$ based on L_1-norm as:

$$\begin{aligned} g\big(\mathbf{d}_{\mathbf{p}(x,y)}, \varepsilon\big) = & \left\| \frac{\mathbf{d}_{\mathbf{p}(x+\varepsilon,y)} - \mathbf{d}_{\mathbf{p}(x,y)}}{\varepsilon} \right\|_1 \\ & + \left\| \frac{\mathbf{d}_{\mathbf{p}(x,y+\varepsilon)} - \mathbf{d}_{\mathbf{p}(x,y)}}{\varepsilon} \right\|_1, \end{aligned} \tag{12}$$

with ε set to be 1 in our work. The gradient loss stimulates the network to compare estimated depths with adjacent pixels. In textureless regions associated with local ambiguities, estimated depth values in these regions are inaccurate. The gradient loss helps to increase smoothness within homogeneous regions [54]. The erroneous estimates can be adjusted by adapting to the surrounding depth variance. However, larger errors are produced if the same gradient loss is applied without adaptive attention weights. This is due to reducing the penalty from the other loss in textured regions (see Subsection IV-D2).

The second part of the total loss, $\mathcal{L}_d\big(\widetilde{\mathbf{d}}_{\mathbf{p}}, \mathbf{d}_{\mathbf{p}}^{gt}\big)$, is the conventional smooth L_1-norm loss [55], which directly optimizes the absolute depth error between the ground-truth depth and the estimated depth, and is defined as follows:

$$\mathcal{L}_d\big(\widetilde{\mathbf{d}}_{\mathbf{p}}, \mathbf{d}_{\mathbf{p}}^{gt}\big) = \Big\| \widetilde{\mathbf{d}}_{\mathbf{p}(x,y)} - \mathbf{d}_{\mathbf{p}(x,y)}^{gt} \Big\|_{s1}. \tag{13}$$

The proposed attention-balanced loss is similar to the popular self-supervised learning, in which the network learns subtle attention weights for different pixels to bring the smallest depth error.

IV. Experiments

A. Datasets and Evaluation Metrics

In our experiments, we use DTU [26] and BlendedMVS [29] datasets for training and evaluate model performance on DTU and Tanks & Temples [27] benchmarks.

DTU [26] is an indoor MVS dataset with a fixed camera trajectory, containing 124 scenes in total. The dataset is divided into the training, validation, and testing sets, comprising 79, 23 and 22 scenes, respectively, [56]. The original image resolution is 1600 × 1200. The aligned ground-truth depth maps, masks and camera parameters are provided in [15]. This dataset can be employed for quantitative analysis of depth estimates and reconstructed point clouds. The Mean Absolute depth Error (MAE) is adopted to evaluate the accuracy of estimated depth maps, while the accuracy and completeness of the distance metric are used to evaluate reconstructed point clouds.

Tanks & Temples [27] provides both outdoor and indoor scenes in realistic illumination conditions with a wide range of scales. It includes intermediate and advanced sets with two resolutions of 2048 × 1080 and 1920 × 1080. The official website provides the images and corresponding camera parameters. However, as no ground-truth depths are provided, it can only be used to analyze reconstructed point clouds quantitatively. For the evaluation metric, the percentages of points with precision and recall for a 2 mm threshold are first measured. Then the harmonic mean of two terms is calculated and denoted as F-score.

BlendedMVS [29] is a recently proposed synthetic dataset containing 113 scenes, including small objects, outdoor sculptures, architectures and larger-scale buildings. The images and ground-truth depth maps are rendered from textured meshes using Blender software. The image resolution is 768 × 576. However, as no ground-truth 3D point clouds are provided, we only use this dataset for model fine-tuning without evaluation.

B. Implementation Details

Following the common practice, the model is first trained on the DTU training set, evaluated on the DTU testing set, and then fine-tuned on the BlendedMVS for generalizability evaluation on the Tanks & Temples benchmark. We set the input resolution to 640×512 and the number of input images to 5 ($N = 4$). The extracted feature channels and the number of depth candidates are set to 2^{k+2} for stage k. The constructed cascade feature volumes have the sizes of $\frac{W}{8} \times \frac{H}{8} \times 64$, $\frac{W}{4} \times \frac{H}{4} \times 32$, $\frac{W}{2} \times \frac{H}{2} \times 16$, $W \times H \times 8$ at stage k from 4 to 1. Since rendered ground-truth depths are bounded with masks and we need to calculate gradient maps from ground-truth depths as supervision, the calculated gradient maps are influenced by the holes of the ground-truth depths, especially for the holes situated in the middle of objects. To address this problem, we pre-train our model without gradient supervision for 10 epochs to get initial depth estimates without masks for training images. Then, we replenish holes in the ground-truth depths with our estimates for the formal training of 15 epochs. Before the evaluation on Tanks & Temples, the model is fine-tuned on BlendedMVS for 10 epochs. The input resolution is set to 768 × 576 and the number of input images to 7 ($N = 6$). For the other experiment setups, we follow the baseline [16]. Adam optimizer is adopted for network training with an initial learning rate of 0.001, and the learning rate is halved after 10, 12, and 14 epochs. The batch size is set to 4, and the model is trained on 2 Nvidia GTX 3090 GPUs. For all the experiments, UCSNet [16] is adopted as the baseline of the depth estimation network. In the ablation study of Subsection IV-D, we additionally employ CasMVSNet [41] to confirm the versatility and effectiveness of our approach. For benchmark evaluations, on the DTU dataset, the input image resolution is set to be 1600 × 1152, and the input number of views is set to 5. On the Tanks & Temples dataset, the input image resolutions are set to 2048 × 1024 and 1920 × 1204, while the input number of views is set to 7. After estimating multi-view depth maps, we use the fusion method [43] to generate final point clouds.

C. Comparisons With State-of-the-Arts

1) Evaluation on DTU benchmark: Table I compares quantitatively the performance of our GA-MVS and 16 existing state-of-the arts, including transformer-based models [17],

TABLE I: Evaluation on DTU benchmark [26]

Methods	Year	Acc.(mm)↓	Comp.(mm)↓	Average (mm)↓
AttMVS [17]	2020	0.383	0.329	0.356
CasMVSNet [41]	2020	0.325	0.385	0.355
PatchmatchNet [43]	2020	0.427	0.277	0.352
UCSNet [16]	2020	0.338	0.349	0.344
AA-RMVSNet [33]	2021	0.376	0.339	0.357
EPP-MVSNet [57]	2021	0.413	0.296	0.355
MVSTR [37]	2021	0.356	0.295	0.326
AACVP-MVSNet [35]	2021	0.357	0.326	0.341
NP-CVP-MVS [46]	2022	0.356	0.275	0.315
CDS-MVSNet [32]	2022	0.351	0.280	0.315
IterMVS [58]	2022	0.373	0.354	0.363
GBINet [44]	2022	0.327	**0.268**	**0.298**
RayMVSNet [18]	2022	0.341	0.319	0.330
MVSTER [36]	2022	0.350	0.276	0.313
TransMVSNet [34]	2022	0.321	0.289	0.305
UGNet [8]	2022	0.334	0.330	0.332
GA-MVS (Ours)	2023	**0.317**	0.302	0.309

[18], [34]–[37], [57], iterative-based models [43], [44], [58] and other advanced models [8], [16], [32], [33], [41], [46], in terms of accuracy and completeness as well as the average of both metrics. In the table, the boldfaced value indicates the best performance, and the underlined value means the second best. Our model achieves the best accuracy, and it is the third best in terms of the average of accuracy and completeness. This shows that our model is very competitive. Qualitatively, our model estimates high-quality depth maps, particularly for intractable regions with shadows or little textures, as shown in Fig. 1 given in the introduction section. This is attributed to our geometry-enhanced features and adaptive constraint strategy, which tackle the ill-posed matching issues by cross-view consistent feature representation and adaptive multi-view matching measurement. First, our geometry-enhanced feature is insensitive to illumination variation, consistent across different views, favorable for robust feature matching and accurate depth estimation. Second, the proposed attention-balanced loss promotes our model to learn the patterns of depth estimation in textureless and general textured regions. Consequently, the estimated depths vary with less fluctuation in textureless regions and lean towards matching results in general textured regions. We visualize some reconstructed results on the DTU testing set [26] by our method in Fig. 5, which shows that our reconstructions are dense and accurate.

TABLE II: GPU memory and runtime comparison

Methods	Image Size	Memory (MB) ↓	Running time (s) ↓
UCSNet [16]	640 × 512	2873	0.342
GA-MVS (Ours)	640 × 512	3321	0.387

The performance data are collected with one batch size on an NVIDIA GTX 3090 GPU card.

Table II compares the GPU memory and runtime of our model with those of the baseline UCSNet model. It can be seen that with the additional GAM, FFM and the attention auxiliary branch, our model only has a marginal increase in memory consumption and runtime, compared with the baseline, while its depth estimation improvements over the latter are very considerable on both the DTU and Tanks & Temples datasets.

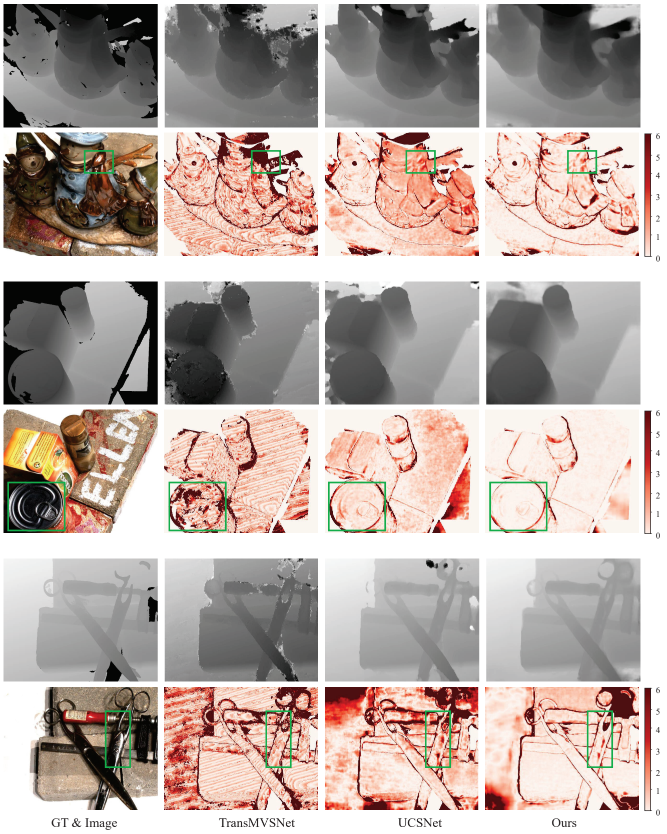


Fig. 6: Visual comparison of estimated depth maps utilizing GA-MVS, TransMVSNet [34] and baseline UCSNet [16]. The green boxes in the observed images indicate regions with specular reflection.

We further verify the depth estimation performance of our GA-MVS in the presence of reflections and specular reflec-

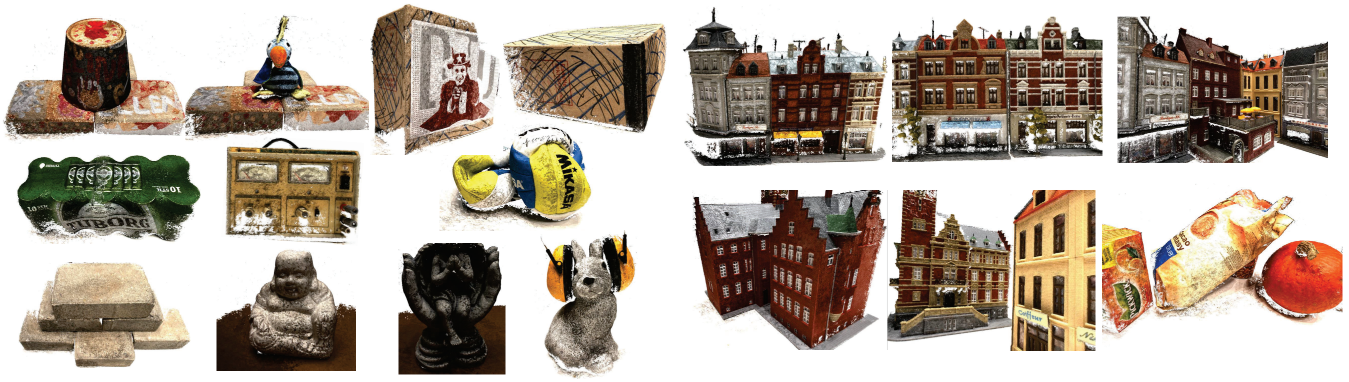


Fig. 5: Reconstruction results on DTU's testing set by our proposed approach

TABLE III: Quantitative performance on scenes with different matching difficulty

	Rich-textured Set	Textureless Set	Photometric-varied Set
UCSNet [16]	0.326	0.371	0.328
Ours	0.311	0.313	0.301
Improvement Ratio	4.4%	15.8%	8.4%

The evaluation metric for the first two rows is the averaged error of reconstructed point clouds, with mm as unit.

tions on the DTU dataset [26]. We select scenes with specular reflection in the DTU validation set to evaluate the depth estimation performance of our method qualitatively. Fig. 6 compares the results of our method with the baseline UCSNet [16] and TransMVSNet [34]. Clearly, our method outperforms these two counterparts with lower estimation biases. Crucially, our method is more general and robust in challenging cases, including textureless and reflection regions, owing to its cross-view consistent feature representation and attentive learning framework. Quantitatively, We create three subsets focusing on scenes with rich-textured, textureless, and photometric-varied situations, from the DTU testing set. The rich-textured

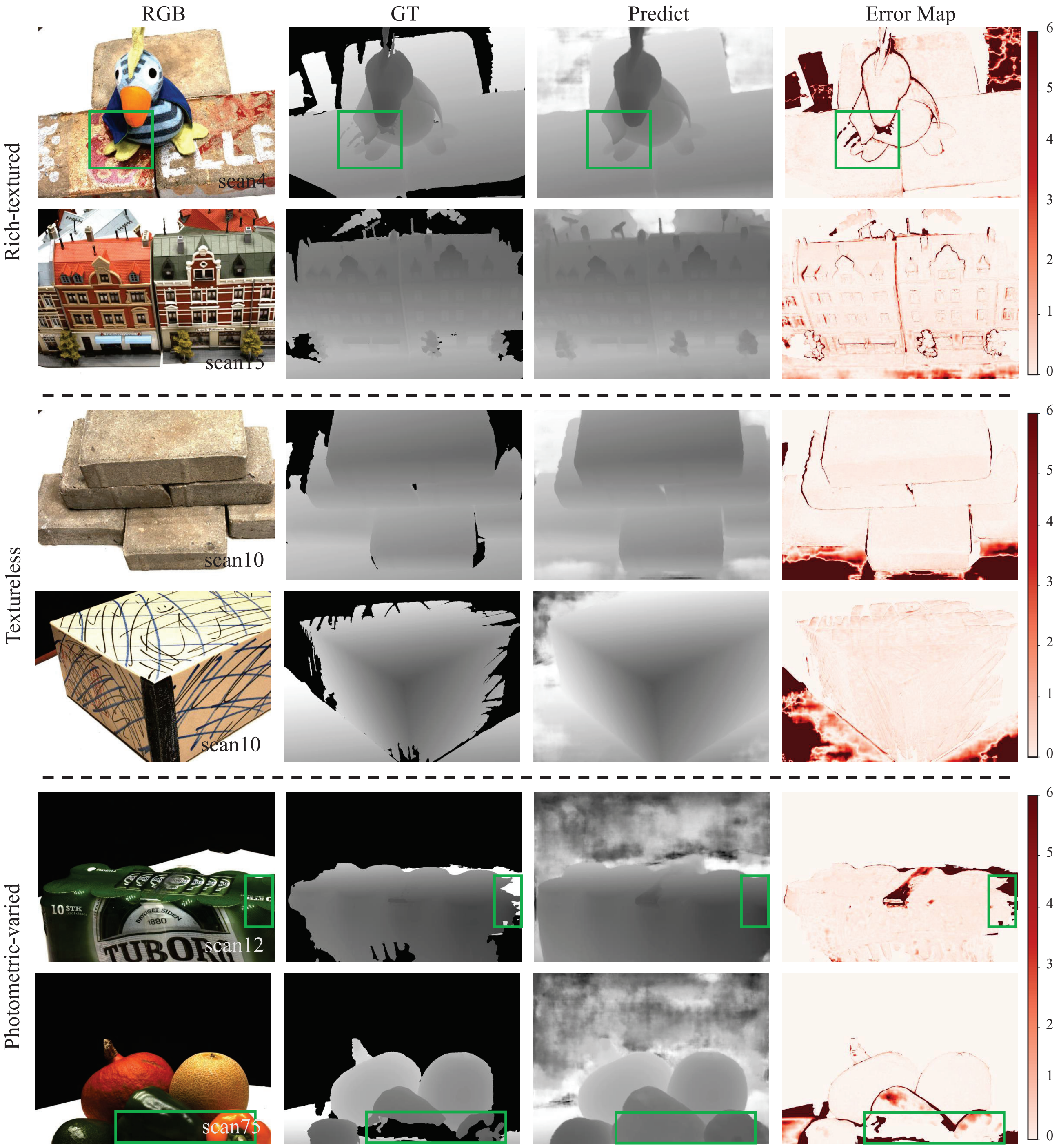


Fig. 7: More depth estimation results on DTU benchmark by our method. Because the ground-truth depth maps are obtained from 3D mesh projection, there exist biased ground-truth depths in green boxes due to the imperfection of the 3D meshes.

set[1] includes the scenes with general rich textures, such as buildings and plush toys. The textureless set[2] focuses on the scenes with large textureless region. The photometric-varied set[3] focuses on the scenes whose appearance is significantly different in the reference and source images due to the influences caused by specular reflection and shadow. We compare the reconstruction results of our model and the baseline model in Table III. The results demonstrate that the proposed method is capable of improving the reconstruction quality of scenes with varied matching difficulty, especially for challenging textureless and photometric-varied situations, achieving the improvement ratios of 15.8% and 8.4%, respectively, over the baseline. Fig. 7 provides more depth estimation results of our method on different subsets.

2) Generalization on Tanks & Temples benchmark: To verify the generalization capability of our method, we evaluate the performance on the Tanks & Temples benchmark using the model fine-tuned on BlendedMVS. The corresponding quantitative results of the reconstructed point clouds on both intermediate and advanced sets are shown in Table IV, in comparison with three traditional MVS methods and 13 learning-based MVS methods. We observe that the proposed GA-MVS obtains competitive F-scores on the both sets, and it achieves the best average F-score over the advanced set, indicating the strong generalization of our model. Fig. 8 compares the respective error maps of partial scenes obtained by the baseline UCSNet [16], the two best performing existing models UGNet [8] and TransMVSNet [34] as well as our GA-MVS. It can be seen that our reconstruction accuracy is particularly good in textureless regions while high accuracy is maintained in other areas, indicating robust depth estimation performance in regions with varying matching difficulty. The evaluated scenes in Tanks & DTU datasets have different depth ranges, and various view changes modes. Our model can well adapt to these entirely different scenes, owing to its cross-view consistent feature representation and adaptive multi-view matching measurement. Fig. 9 illustrates the reconstructed point clouds on the Tanks & Temples benchmark achieved by our method. Our reconstruction results are complete and with rich details, demonstrating the robustness of GA-MVS. Fig. 10 offers additional qualitative results of GA-MVS on Tanks &

[1] scan4,scan9,scan15.scan23,scan29,scan32,scan49,scan62,scan75

[2] scan10,scan11,scan12,scan13,scan33,scan34,scan48

[3] scan1,scan24,scan77,scan110,scan114,scan118

TABLE IV: Evaluation on Tanks & Temples benchmark [27]

Method	Year	Intermediate Set ↑									Advanced Set ↑						
		Mean	Fam.	Fran.	Hor.	Lig.	M60	Pan.	Pla.	Tra.	Mean	Aud.	Bal.	Cou.	Mus.	Pal.	Tem.
OpenMVS* [13]	2015	55.11	71.69	51.12	42.76	58.98	54.72	56.17	59.77	45.69	34.43	24.49	38.39	38.21	48.48	27.25	31.79
COLMAP* [14]	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
ACMH* [59]	2019	54.82	69.99	49.45	45.12	59.04	52.64	52.37	58.34	51.61	33.73	21.69	32.56	**40.62**	47.27	24.04	36.17
PatchmatchNet [43]	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
UCSNet [16]	2020	54.03	76.09	53.16	43.03	54.00	55.60	51.49	57.38	47.89	-	-	-	-	-	-	-
CasMVSNet [41]	2020	56.84	76.37	58.45	46.26	55.81	56.11	54.06	58.18	49.51	31.12	19.81	38.46	29.10	43.87	27.36	28.11
AttMVS [17]	2020	60.05	73.90	62.58	44.08	64.88	56.08	59.39	**64.42**	56.06	31.93	15.96	27.71	37.99	**52.01**	29.07	28.84
Vis-MVSNet [60]	2020	60.03	77.40	60.23	47.07	63.44	62.21	57.28	60.54	52.07	33.78	20.79	38.77	32.45	44.20	28.73	37.70
AA-RMVSNet [33]	2021	61.51	77.77	59.53	51.53	64.02	64.05	59.47	60.85	54.90	33.53	20.96	40.15	32.05	46.01	29.28	32.71
AACVP-MVSNet [35]	2021	58.39	78.71	57.85	50.34	52.76	59.73	54.81	57.98	54.94	-	-	-	-	-	-	-
NP-CVP-MVS [46]	2022	59.64	78.93	64.09	51.82	59.42	58.39	55.71	56.07	52.71	-	-	-	-	-	-	-
CDS-MVSNet [32]	2022	60.82	78.17	61.74	53.12	60.25	61.91	58.45	62.35	50.58	-	-	-	-	-	-	-
RayMVSNet [18]	2022	59.49	78.56	61.96	45.48	57.58	61.01	59.76	59.20	52.32	-	-	-	-	-	-	-
MVSTER [36]	2022	60.92	80.21	63.51	52.30	61.38	61.47	58.16	58.98	51.38	37.53	**26.68**	42.14	35.65	49.37	32.16	**39.19**
TransMVSNet [34]	2022	63.52	**80.92**	65.83	**56.94**	62.54	63.06	60.00	60.2	**58.67**	37.00	24.84	**44.59**	34.77	46.49	**34.69**	36.62
UGNet [8]	2022	**64.12**	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
GA-MVS (Ours)	2023	63.30	79.71	**67.67**	54.75	61.25	64.54	**62.04**	59.67	56.75	**38.04**	26.32	42.97	38.31	51.20	31.62	37.84

* indicates traditional MVS methods, the others are learning-based MVS methods. The evaluation metric is the F-score using percentage metric, which considers both accuracy and completeness of final reconstructed point cloud results. All the values, including ours, are available in the website [61].

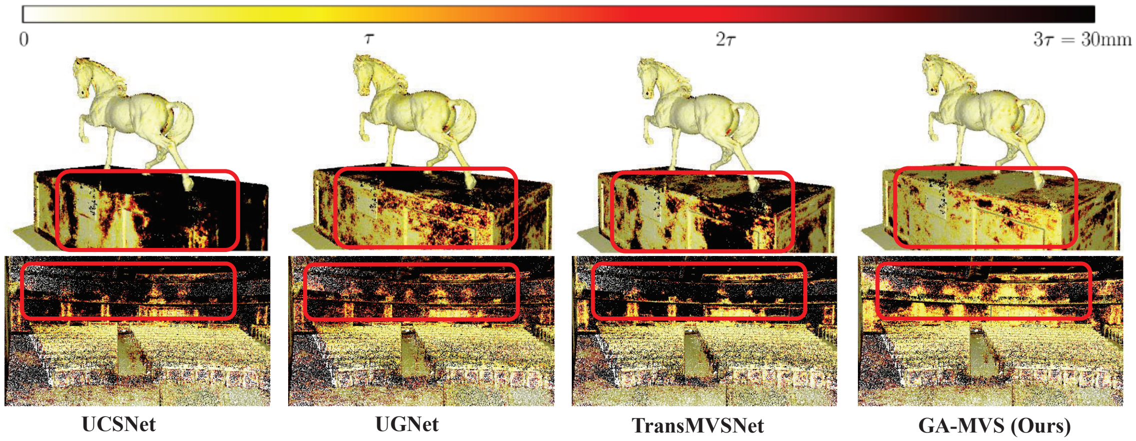


Fig. 8: Error visualization of Horse and Auditorium scenes in the Tanks & Temples benchmark [27]. We exemplify the errors of baseline UCSNet [16], two best performing existing models UGNet [8] and TransMVSNet [34], and ours, computed based on ground-truth point clouds. Darker points indicate bigger errors of reconstructions.

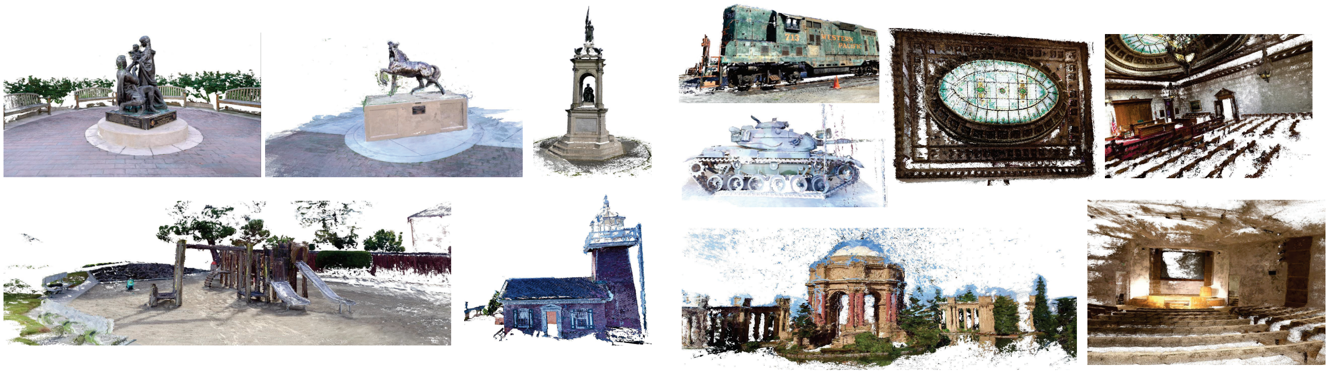


Fig. 9: Reconstruction results on Tanks & Temples dataset by our proposed approach

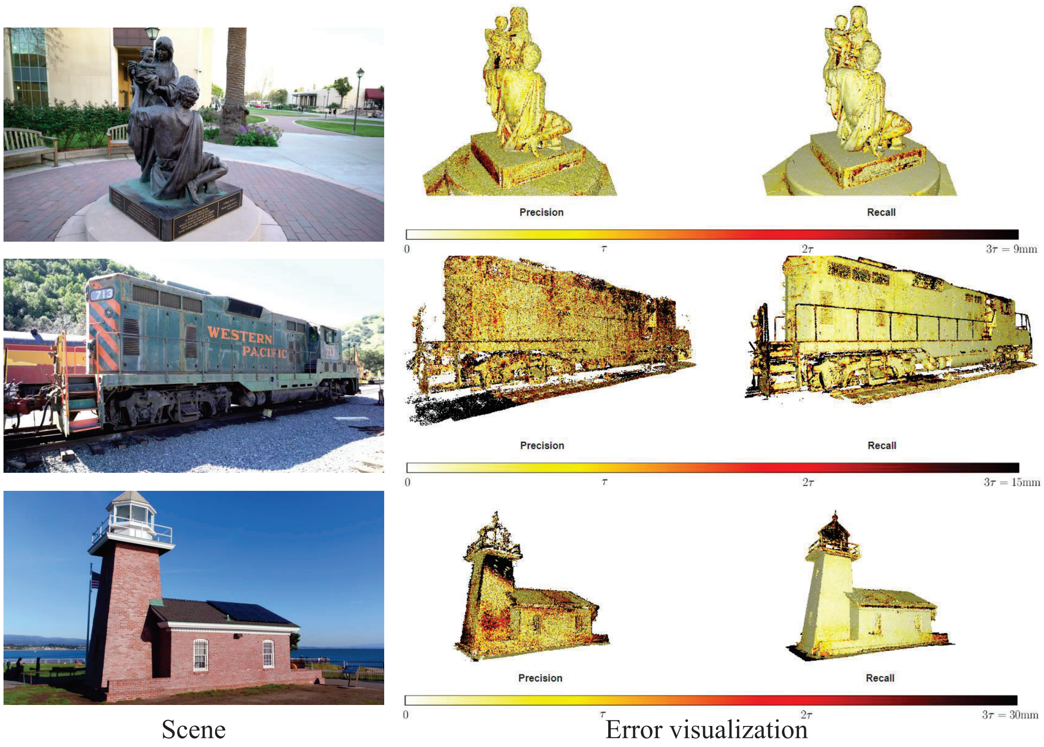


Fig. 10: More qualitative results of our method on Tanks & Temples benchmark (family, train, and lighthouse scenes).

Temples, in terms of reconstructed point clouds. It can be seen that our GA-MVS achieves high-quality reconstruction in various scenes, demonstrating its generalization capability.

D. Ablation Study

In this subsection, we analyze the effects of the GA-MVS components. We adopt two popular cascaded MVS methods, UCSNet [16] and CasMVSNet [41], as the baselines to analyze the impacts of the proposed components. The both baselines use the variance-based fusing metric to construct multi-view cost volumes and use 3D-CNN for cost volume regularization. Differently, UCSNet estimates the variance-based confidence intervals centering on the previous estimates to construct the cascade cost volumes, and CasMVSNet progressively narrows the depth range. UCSNet adopts 2D-UNet for feature extraction, while CasMVSNet adopts 2D-FPN. We modify the two baselines for a fair comparison to construct the four scale cost volumes of the same sizes as ours. Accordingly, the feature extractors for the both baselines are modified to output four scale features, while preserving their UNet and FPN structures, respectively. Table V provides the detailed network configurations. Other training settings are kept the same as our implementation.

1) Geometry-enhanced Features: The results in Table VI indicate that geometry-enhanced features improve the accuracy and completeness of reconstructions on the both baselines. This improvement can be attributed to the two aspects: i) Benefited from applying the feature loss of the GAM, the extracted geometric features are free from view-dependent photometric effects, thus enhancing the robustness of multi-view matching; ii) Discriminative texture and stable geometric features are effectively integrated via the proposed FFM.

1.1) Discussion on GAM and FFM: To deeply investigate the influences of the aforementioned two aspects, we further analyze the proposed GAM and FFM, which are utilized to obtain geometry-enhanced features. For the texture feature

TABLE V: Network Architectures of UNet and FPN Texture Features (adopted in Tables VI and VII of Ablation Study)

Input	Description	Output	Output Shape	Input	Description	Output	Output Shape
UNet-Structured Feature Extractor ($I \rightarrow F'_k$)				Ours: **FPN-Structured** Feature Extractor ($I \rightarrow F'_k$)			
I	Conv 3×3 Unit	X0	$H \times W \times 8$	I	Conv3 $\times$ 3 Unit	X0	$H \times W \times 8$
X0	Conv 3×3 Unit	X1	$H \times W \times 8$	X0	Conv 3×3 Unit	X1	$H \times W \times 8$
X1	Conv 5×5 Unit	X2	$H/2 \times W/2 \times 16$	X1	Conv 5×5 Unit	X2	$H/2 \times W/2 \times 16$
X2	Conv 3×3 Unit	X3	$H/2 \times W/2 \times 16$	X2	Conv 3×3 Unit	X3	$H/2 \times W/2 \times 16$
X3	Conv 3×3 Unit	X4	$H/2 \times W/2 \times 16$	X3	Conv 3×3 Unit	X4	$H/2 \times W/2 \times 16$
X4	Conv 5×5 Unit	X5	$H/4 \times W/4 \times 32$	X4	Conv 5×5 Unit	X5	$H/4 \times W/4 \times 32$
X5	Conv 3×3 Unit	X6	$H/4 \times W/4 \times 32$	X5	Conv 3×3 Unit	X6	$H/4 \times W/4 \times 32$
X6	Conv 3×3 Unit	X7	$H/4 \times W/4 \times 32$	X6	Conv 3×3 Unit	X7	$H/4 \times W/4 \times 32$
X7	Conv 5×5 Unit	X8	$H/8 \times W/8 \times 64$	X7	Conv 5×5 Unit	X8	$H/8 \times W/8 \times 64$
X8	Conv 3×3 Unit	X9	$H/8 \times W/8 \times 64$	X8	Conv 3×3 Unit	X9	$H/8 \times W/8 \times 64$
X9	Conv 3×3 Unit	X10	$H/8 \times W/8 \times 64$	X9	Conv 3×3 Unit	X10	$H/8 \times W/8 \times 64$
X10	Conv 1×1	F'_4	$H/8 \times W/8 \times 64$	X10	Conv 1×1	F'_4	$H/8 \times W/8 \times 64$
X10	TransConv 3×3 Unit	X11	$H/4 \times W/4 \times 32$	X10,X7	BI (X10)+Conv 1×1 (X7)	X11	$H/4 \times W/4 \times 64$
X11,X7	Concat+Conv 3×3 Unit	X12	$H/4 \times W/4 \times 32$	X11	Conv 1×1	F'_3	$H/4 \times W/4 \times 32$
X12	Conv 1×1	F'_3	$H/4 \times W/4 \times 32$	X11,X4	BI (X10)+Conv 1×1 (X4)	X12	$H/2 \times W/2 \times 64$
X12	TransConv 3×3 Unit	X13	$H/2 \times W/2 \times 16$	X12	Conv 1×1	F'_2	$H/2 \times W/2 \times 16$
X13,X4	Concat+Conv 3×3 Unit	X14	$H/2 \times W/2 \times 16$	X12,X1	BI (X12)+Conv 1×1 (X1)	X13	$H \times W \times 64$
X14	Conv 1×1	F'_2	$H/2 \times W/2 \times 16$	X13	Conv 1×1	F'_1	$H \times W \times 8$
X14	TransConv 3×3 Unit	X15	$H \times W \times 8$				
X15,X1	Concat+Conv 3×3 Unit	X16	$H \times W \times 8$				
X16	Conv 1×1	F'_1	$H \times W \times 8$				

Conv and TransConv denote 2D convolution and 2D transposed convolution (also known as deconvolution). Each convolutional unit comprises a 2D convolution layer, a BN (batch normalization) layer, and a ReLU layer. Conv1x1 applies only a single 2D convolution layer. BI represents bilinear interpolation. F'_k marked in red present the output multi-scale texture features.

TABLE VI: Ablated results of employing different components on DTU testing set

Model	Step	Acc.(mm) ↓	Comp.(mm)↓	Average(mm)↓	R.(%)↑
UCSNet [16]	Baseline	0.328	0.347	0.338	-
	Geo. Features + Baseline	0.323	0.332	0.328	2.96
	Geo. Features + Baseline + Att. Loss	**0.321**	**0.319**	**0.320**	2.44
CasMVSNet [41]	Baseline	0.322	0.379	0.351	-
	Geo. Features + Baseline	0.317	0.358	0.338	3.85
	Geo. Features + Baseline + Att. Loss	**0.315**	**0.346**	**0.331**	2.07

Geo. Features indicate geometry-enhanced features detailed in Subsection III-A, and Att. Loss indicates attention-balanced loss detailed in Subsection III-D. R. column indicates the improvement ratio, in terms of overall quality, by adding one more component.

extraction, we compare the UNet and FPN structures. From the results (the 1st and 4th rows) of Table VII, we notice that FPN is better than UNet. We infer that the different manners of multi-stage feature integration plays an essential role. The FPN adopts up-sampling and element-wise addition operations for integration, which retain more detailed texture information and is beneficial for obtaining discriminative matching results. The UNet adopts concatenation and deconvolution layers, and after the transformation of convolution and nonlinear activation, the original shallow features may not be well preserved, which may harm accurate matching measure. Further considering the algorithm efficiency, we adopt the FPN for texture feature extraction in our method.

To verify the effectiveness of extracted geometric features, we retain the initial fusion step and the final 1×1 convolution in the FFM but remove the channel attention (CA). As shown in the 2nd and 5th rows of Table VII, the models with (GAM+FFM) w/o CA perform better overall than the respective baseline models (the 1st and 4th rows). This confirms that geometric features extracted by GAM are nontrivial for accurate depth estimation. To validate the effectiveness of the feature integration operation, we add CA to retrain the models with the complete FFM. As shown in the 3rd and 6th rows of Table VII, the models with (GAM + FFM) perform better than their respective models without CA. It can also be seen that our proposed method achieves the best overall result, indicating the effective contribution of CA and the proposed FFM for final depth estimates.

TABLE VII: Ablated results of GAM and FFM with different components on DTU testing set.

Model	Acc.(mm) ↓	Comp.(mm)↓	Average(mm)↓	R.(%)↑
UNet	0.328	0.347	0.338	-
UNet + GAM + FFM w/o CA	0.325	0.346	0.336	0.59
UNet + GAM + FFM	**0.323**	**0.332**	**0.328**	2.38
FPN	0.324	0.341	0.333	-
FPN + GAM + FFM w/o CA	**0.321**	0.327	0.324	2.70
FPN + GAM + FFM (Ours)	0.322	**0.301**	**0.312**	3.70

In our GAM, we incorporate the lowest-level feature F'_{1i} with the highest-level feature F'_{4i} to form the one-channel geometric feature. We also test the effects of various combinations of two features for the GAM to access geometric features. The results presented in Table VIII indicate that the our combination of $F'_{1i} + F'_{4i}$ achieves the highest overall quality for our 3D reconstruction task, while $F'_{1i} + F'_{2i}$ achieves the worst

TABLE VIII: Quantitative results for varied inputs of GAM on DTU testing set

Model	Acc.(mm)↓	Comp.(mm)↓	Average(mm)↓
$F'_{1i} + F'_{2i}$	0.344	0.328	0.336
$F'_{1i} + F'_{3i}$	0.343	0.311	0.327
$F'_{1i} + F'_{4i}$ (Ours)	**0.322**	**0.301**	**0.312**
$F'_{2i} + F'_{3i}$	0.339	0.317	0.328
$F'_{2i} + F'_{4i}$	0.326	0.306	0.316
$F'_{3i} + F'_{4i}$	0.348	0.304	0.326

$\{F'_{ki}\}_{k=1}^{4}$ represents the feature map of spatial resolution $\frac{W}{2^{k-1}} \times \frac{H}{2^{k-1}}$

overall quality. We infer this is because the texture details contained in low-level features inevitably cause interference for stable geometric feature learning, which is detrimental to the effectiveness of GAM. For the other combinations, when the input features are deeper, the overall accuracy of reconstruction gains slightly, which indicates that the high-level feature may influence more on the geometric feature learning procedure. We observe that the overall reconstruction quality varies in a relatively small range of 0.312 to 0.336 for different input combinations.

1.2) Visualization of Geometry-enhanced Features: We compare the geometry-enhanced features with the FPN-extracted texture features in Fig. 11. Specifically, both features are averaged and normalized along the channel dimension for visualization. It can be seen that feature differences caused by illumination are effectively alleviated by our geometry-enhanced features, thus ensuring the robustness of multi-view matching.

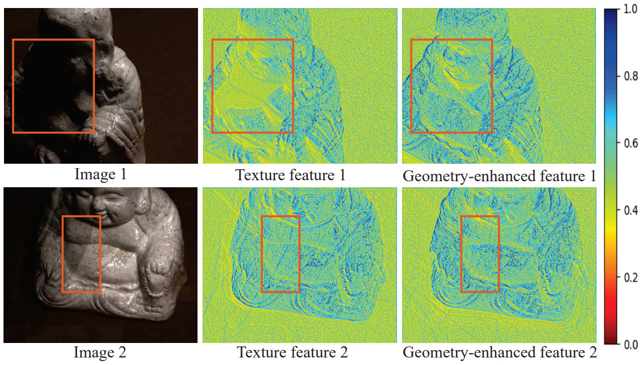


Image 1 Texture feature 1 Geometry-enhanced feature 1

Image 2 Texture feature 2 Geometry-enhanced feature 2

Fig. 11: Comparison of initial texture features and geometry-enhanced features corresponding to two input images. Our geometry-enhanced features can effectively restrain view-dependent photometric effects in shadow and shadow boundary regions.

2) Attention-balanced Loss: For final 3D reconstructions, the effectiveness of the proposed attention-balanced loss is confirmed by the ablated results shown in Table VI (the second vs. third rows, and the fifth vs. sixth rows). It can be seen that the two baseline structures with the attention-balanced loss yield the gains of 2.44% and 2.07%, respectively, in terms of overall quality, indicating that the proposed attention-balanced loss is a general component that can be combined with other depth estimation networks to boost their performances.

For depth estimation, we evaluate the effectiveness of the attention-balanced loss by contrasting it with the fixed combined losses and the conventional smooth L_1 loss on the DTU validation set. For the versions without attention-balanced loss, we extract geometry-enhanced features with the FPN version and utilize the four-stage UCSNet as the depth estimation network for model training. For the evaluation metrics, the MAE in millimeters is adopted to quantify the average depth error of all pixels, while 2 mm (%) and 8 mm (%) criteria, which are the percentages of pixels with absolute depth errors smaller than thresholds 2 mm and 8 mm, respectively, indicate the capacity of a model to handle the challenging situation of textureless areas with larger errors. Table IX provides a summary of the evaluation results.

TABLE IX: Comparison of evaluation results obtained with various losses on DTU validation set

Model	MAE(mm)↓	<2mm(%)↑	<8mm(%)↑
Depth loss only	5.52	73.74%	88.39%
Gradient loss only	89.96	0.19%	1.34%
Depth + Gradient	6.14	71.25%	90.62%
Attention-balanced loss	**4.63**	**75.74%**	**91.39%**

The numbers denote the MAE of all valid pixels in the DTU validation set. The percentages in the table denote the ratio of pixels with depth errors less than 2 mm or 8 mm.

2.1) Discussion on Attention-balanced Loss: As can be seen from Table IX, the fixed combination of gradient and depth loss outperforms the version with depth loss alone in terms of the 8 mm (%) metric. This is because the gradient loss is activated when the model output error-estimated depth fluctuates, which mainly occurs in textureless areas, associated with matching ambiguities. Since large depth estimation errors mainly exist in textureless regions, the addition of gradient loss brings better constraints on these areas. Nevertheless, we observe that with the fixed combined losses, the MAE metric is degraded compared to the depth loss alone, and the network cannot converge solely with the gradient loss. This can be explained by the fact that the gradient loss ignores the predicted depth value of pixels but only offers the difference between neighbors, leading to a dilution of the penalty in general textured areas.

For the proposed attention-balanced loss, we witness the lowest MAE and the highest average percentages of pixels within the thresholds of 2 mm and 8 mm, indicating fewer pixels are estimated with larger errors. Our GA-MVS can learn adaptive attention weights for different pixels. Specifically, higher attentions are learned for pixels in textureless areas to intensify the gradient loss penalty, thus improving depth estimation precision in such regions. For pixels in textured areas, our GA-MVS learns to lower the gradient loss penalty, thus avoiding unfavorable impact on depth estimates.

2.2) Visualization of Attention-balanced Loss: Fig. 12 provides visual examples of the attention maps for scenes with no prominent textures. Visualizing the learned attention maps can give us more insights into how the model works better in textureless regions. It can be observed from Fig. 12 that regions with light colors have higher attention weights. Accordingly, the error maps of our method exhibit lower depth errors in these areas compared to the baseline model [16]. Without attention maps, the estimated depths in textureless regions usually have relatively large biases with mottled red, caused by matching ambiguity. The attention map learned from RGB+edge channels helps the model to recognize this situation, resulting in higher weights of gradient loss (lighter color of attention maps). Higher penalties on gradient loss influence the depth estimation model to produce depth with less fluctuation. Meanwhile, the depth loss term promotes global correctness. Hence, inaccuracies in textureless areas are rectified by minimizing the attention-balanced loss. In contrast,

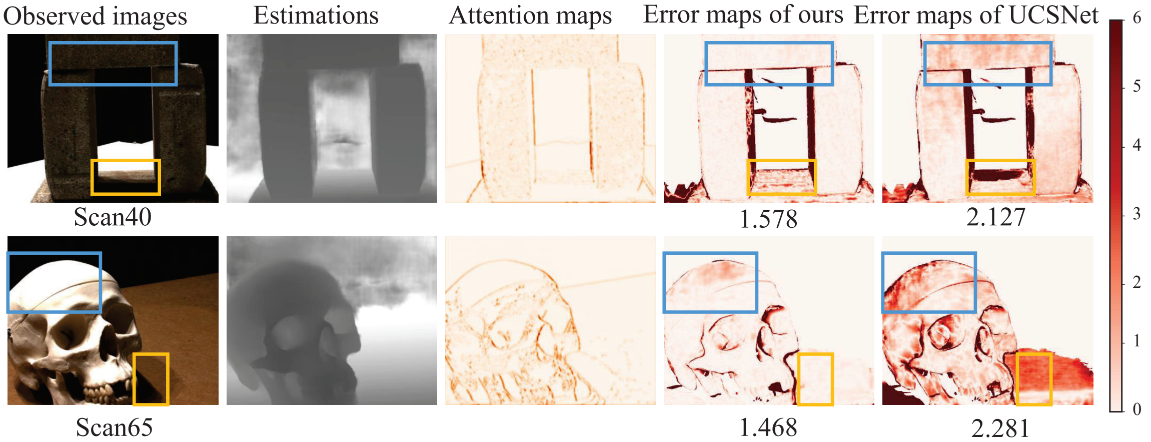


Fig. 12: Visual comparisons for images of scan40 and scan65 scenes. Blue boxes indicate textureless regions, while yellow boxes show shadow regions. Numbers under depth error maps indicate their MAE.

the baseline model cannot distinguish between reliable and unreliable matching results, leading to large errors in challenging textureless regions.

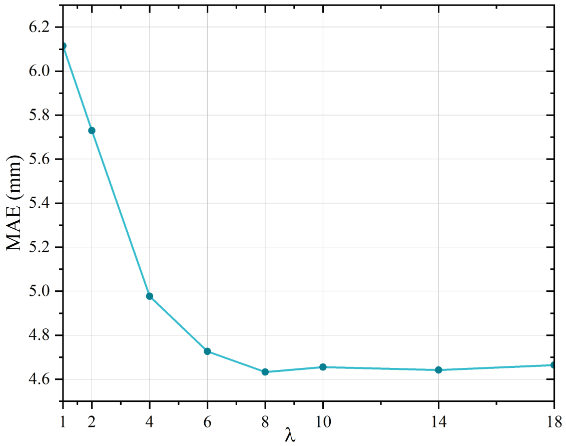


Fig. 13: Ablation study on hyperparameter λ

2.3) Ablation Study of Loss Weight λ: We carry out an ablation study on the hyperparameter λ in Eq. (10). As illustrated in Fig. 13, $\lambda = 8$ is an appropriate weight for the proposed attention-balanced loss.

3) Number of Cascade Stages: Most recent coarse-to-fine MVS models realize depth estimation via three or four stages and encode the input images into three or four-scale feature maps. For example, PatchmatchNet [43] and GBINet [44] adopt four-stage architecture to advance depth map estimation in a coarse-to-fine manner, while UCSNet [16] and UGNet [8] adopt three-stage architecture. The original version of UCSNet constructs three-stage feature volumes for depth estimation. In our GA-MVS, we set the number of cascade stages of the initial multi-scale texture feature set to 4.

3.1) Effects of Number of Stages: We conduct an ablation study on the number of stages for our method. The results are summarized in Table X. We train two versions of three-stage models, GA-MVS$_{3\_1}$ and GA-MVS$_{3\_2}$, with varying settings of the number of depth candidates and the number of feature channels. GA-MVS$_{3\_1}$ shares the same setting as our four-stage model for the last three stages, while GA-MVS$_{3\_2}$ contains the same setting as the original UCSNet version. The comparison results of GA-MVS$_{3\_2}$ vs. GA-MVS$_{3\_1}$ and GA-MVS$_4$ vs. GA-MVS$_{3\_1}$ demonstrate that both increasing the number of depth candidates and the number of stages lead to improved overall accuracy. Our model achieves the best result when both factors are combined.

TABLE X: Reconstruction quality on DTU dataset with different number of stages

Model	Channel Num.	Depth Num.	Acc.(mm)↓	Comp.(mm)↓	Average(mm)↓
GA-MVS$_{3\_1}$	32,16,8	32,16,8	0.323	0.369	0.346
GA-MVS$_{3\_2}$	32,16,8	64,32,8	0.349	0.313	0.331
GA-MVS$_4$ (Ours)	64,32,16,8	64,32,16,8	**0.317**	**0.302**	**0.309**

TABLE XI: Performance comparison of different stages

Method	Resosution	Acc.(mm)↓	Comp.(mm)↓	Average(mm)↓	GPU Mem. (MB)↓	Run-time (s)↓
Our 1st stage	$1/8 \times 1/8$	0.772	0.752	0.762	**1161**	**0.088**
Our 2nd stage	$1/4 \times 1/4$	0.594	0.460	0.527	2290	0.184
Our 3nd stage	$1/2 \times 1/2$	0.348	0.374	0.361	4104	0.390
Our full model	1	**0.317**	**0.302**	**0.309**	6791	0.757

The statistics are collected on the DTU testing set [26] using our model. The original resolution is 1600×1152. The run-time is the sum of the current and previous stages.

3.2) Achievable Performance at Different Stages: To investigate the achievable performance at different stages for our model, we compare multi-stage model performances on DTU benchmark in terms of reconstruction quality, GPU memory, and run time. The statistics are shown in Table XI, and visualization results are shown in Fig. 14. The overall quality is enhanced from 0.762 to 0.309 in a coarse-to-fine manner. Accordingly, the GPU memory increases from

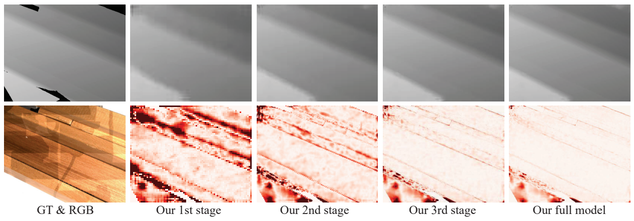


Fig. 14: Depth inference results of each stage. Top-row: Ground truth depth map and intermediate depth inference results. Bottom row: Reference image and error maps of intermediate results.

1161 MB to 6791 MB, and the run-time increases from 0.088 s to 0.757 s. Basically, the number of cascade stages trades off the achievable performance with computational complexity.

V. Conclusions

This paper has developed a Geometry-enhanced Attentive MVS network, called GA-MVS, designed for accurate depth estimation in challenging real-world scenarios with ill-posed matching conditions. Specifically, we have introduced a geometry-enhanced feature extractor that allows for consistent feature representation even in complex lighting conditions, thus enabling robust matching. This novel feature extractor incorporates reliable constraints to facilitate effective feature learning. Additionally, we have proposed an attentive learning framework that enhances depth estimation performance in textureless regions by employing an attention-balanced loss. This adaptive loss encourages the predicted depths to vary with less fluctuation in textureless areas while aligning with matching results in rich textured regions, thereby achieving accurate depth estimation performance in regions with varying texture richness. Experimental results conducted on two benchmarks have verified the effectiveness of our method. The consistently top-performing results validate the superiority and generalizability of our GA-MVS. Furthermore, the introduced attentive learning framework has the potential to be integrated with other regression tasks, such as stereo matching, monocular depth estimation, and image enhancement. As part of our future work, we plan to explore the integration of our modules with other regression tasks to further enhance their performance.

References

[1] C. Yildirim, "Cybersickness during VR gaming undermines game enjoyment: A mediation model," *Displays*, vol. 59, pp. 35–43, Sep. 2019.

[2] H. Kang, J. Ko, H. S. Park, and H. K. Hong, "Effect of outside view on attentiveness in using see-through type augmented reality device," *Displays*, vol. 57, pp. 1–6, Apr. 2019.

[3] C. Gu, *et al.*, "MedUCC: Medium-driven underwater camera calibration for refractive 3-D reconstruction," *IEEE Trans. Systems, Man, and Cybernetics: Systems*, vol. 52, no. 9, pp. 5937–5948, Sep. 2022.

[4] C. Yang, *et al.*, "A comprehensive study of 3-D vision-based robot manipulation," *IEEE Trans. Cybernetics*, vol. 53, no. 3, pp. 1682–1698, Mar. 2023.

[5] F. Rottensteiner, *et al.*, "Results of the ISPRS benchmark on urban object detection and 3D building reconstruction," *ISPRS J. Photogrammetry and Remote Sensing*, vol. 93, pp. 256–271, Jul. 2014.

[6] S. Malihi, *et al.*, "3D building reconstruction using dense photogrammetric point cloud," *Int. Archives Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLI-B3, pp. 71–74, 2016.

[7] G. Bitelli, V. A. Girelli, and A. Lambertini, "Integrated use of remote sensed data and numerical cartography for the generation of 3D city models." *Int. Archives Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLII-2, pp. 97–102, 2018.

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

[9] M. Buyukdemircioglu and S. Kocaman, "Reconstruction and efficient visualization of heterogeneous 3D city models," *Remote Sensing*, vol. 12, no. 13, article no. 2128, pp. 1-26, 2020.

[10] H. Dai, *et al.*, "Adaptive disparity candidates prediction network for efficient real-time stereo matching," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 3099–3110, May 2022.

[11] C. Li, *et al.*, "Hybrid-MVS: Robust multi-view reconstruction with hybrid optimization of visual and depth cues," *IEEE Trans. Circuits and Systems for Video Technology*, early access, May 2023. DOI:10.1109/TCSVT.2023.3276753

[12] C. Bailer, M. Finckh, and H. Lensch, "Scale robust multi view stereo," in *Proc. ECCV 2012* (Florence, Italy), Oct. 7-13, 2012, pp. 398–411.

[13] *Open Multi-View Stereo Reconstruction Library*, 2015. https://github.com/cdcseacave/openMVS

[14] J. L. Schnberger, E. Zheng, J.-M. Frahm, and M. Pollefeys, "Pixelwise view selection for unstructured multi-view stereo," in *Proc. ECCV 2016* (Amsterdam, The Netherlands), Oct. 11-14, 2016, pp. 501–518.

[15] Y. Yao, *et al.*, "MVSNet: Depth inference for unstructured multi-view stereo," in *Proc. ECCV 2018* (Munich, Germany), Sep. 8-14, 2018, pp. 767–783.

[16] S. Cheng, *et al.*, "Deep stereo using adaptive thin volume representation with uncertainty awareness," in *Proc. CVPR 2019*, Jun. 14-19, 2020, pp. 2524–2534.

[17] K. Luo, *et al.*, "Attention-aware multi-view stereo," in *Proc. CVPR 2020*, Jun. 14-19, 2020, pp. 1590–1599.

[18] J. Xi, *et al.*, "RayMVSNet: Learning ray-based 1D implicit fields for accurate multi-view stereo," in *Proc. CVPR 2022* (New Orleans, LA, USA), Jun. 18-24, 2022, pp. 8595–8605.

[19] R. Zhao, *et al.*, "Exploring the point feature relation on point cloud for multi-view stereo," *IEEE Trans. Circuits and Systems for Video Technology*, early access, Apr. 2023. DOI:10.1109/TCSVT.2023.3267457

[20] H. Liu, *et al.*, "Features combined binary descriptor based on voted ring-sampling pattern," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 30, no. 10, pp. 3675–3687, Oct. 2020.

[21] H. Liu, Y. Cong, G. Sun, and Y. Tang, "Robust 3-D object recognition via view-specific constraint," *IEEE Trans. Systems, Man, and Cybernetics: Systems*, vol. 51, no. 11, pp. 7109–7119, Nov. 2021.

[22] B. Fan, *et al.*, "Learning semantic-aware local features for long term visual localization," *IEEE Trans. Image Processing*, vol. 31, pp. 4842–4855, Jul. 2022.

[23] B. Fan, *et al.*, "Seeing through darkness: Visual localization at night via weakly supervised learning of domain invariant features," *IEEE Trans. Multimedia*, vol. 25, pp. 1713–1726, Jun. 2023.

[24] B. Fan, *et al.*, "Deep unsupervised binary descriptor learning through locality consistency and self distinctiveness," *IEEE Trans. Multimedia*, vol. 23, pp. 2770–2781, Aug. 2021.

[25] B. Fan, *et al.*, "A performance evaluation of local features for image-based 3D reconstruction," *IEEE Trans. Image Processing*, vol. 28, no. 10, pp. 4774–4789, Oct. 2019.

[26] R. Jensen, *et al.*, "Large scale multi-view stereopsis evaluation," in *Proc. CVPR 2014* (Columbus, OH, USA), Jun. 23-28, 2014, pp. 406–413.

[27] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun, "Tanks and temples: Benchmarking large-scale scene reconstruction," *ACM Trans. Graphics*, vol. 36, no. 4, article no. 78, pp. 1–13, 2017.

[28] T. Schöps, *et al.*, "A multi-view stereo benchmark with high-resolution images and multi-camera videos," in *Proc. CVPR 2017* (Honolulu, HI, USA), Jul. 21-26, 2017, pp. 2538–2547.

[29] Y. Yao, *et al.*, "BlendedMVS: A large-scale dataset for generalized multi-view stereo networks," in *Proc. CVPR 2020*, Jun. 14-19, 2020, pp. 1790–1799.

[30] B. Liu, H. Yu, and Y. Long, "Local similarity pattern and cost self-reassembling for deep stereo matching networks," in *Proc. AAAI 2022*, Feb. 22-Mar. 1, 2022, pp. 1647–1655.

[31] X. Song, X. Zhao, H. Hu, and L. Fang, "Edgestereo: A context integrated residual pyramid network for stereo matching," in *Proc. ACCV 2018* (Perth, Australia), Dec. 2-6, 2018, pp. 20–35.

[32] K. T. Giang, S. Song, and S. Jo, "Curvature-guided dynamic scale networks for multi-view stereo," in *Proc. ICLR 2022*, Apr. 25-29, 2022, pp. 1–19.

[33] Z. Wei, *et al.*, "AA-RMVSNet: Adaptive aggregation recurrent multi-view stereo network," in *Proc. ICCV 2021* (Montreal, QC, Canada), Oct. 10-17, 2021, pp. 6187–6196.

[34] Y. Ding, *et al.*, "TransMVSNet: Global context-aware multi-view stereo network with transformers," in *Proc. CVPR 2022* (New Orleans, LA, USA), Jun. 19-24, 2022, pp. 8585–8594.

[35] A. Yu, *et al.*, "Attention aware cost volume pyramid based multi-view

stereo network for 3D reconstruction," *ISPRS J. Photogrammetry and Remote Sensing*, vol 175, pp. 448–460, May 2021.

[36] X. Wang, *et al.*, "MVSTER: Epipolar transformer for efficient multi-view stereo," in *Proc ECCV 2022* (Tel Aviv, Israel), Oct. 23-27, 2022, pp. 573–591.

[37] J. Zhu, *et al.*, "Multi-view stereo with transformer," arXiv:2112.00336v1, Dec. 2021, pp. 1–10.

[38] S. M. Seitz, *et al.*, "A comparison and evaluation of multi-view stereo reconstruction algorithms," in *Proc. CVPR 2006* (New York, NY, USA), Jun. 17-22, 2006, pp. 1–8.

[39] Y. Yao, *et al.*, "Recurrent MVSNet for high-resolution multi-view stereo depth inference," in *Proc. CVPR 2019* (Long Beach, CA, USA), Jun. 16-20, 2019, pp. 5525–5534.

[40] J. Yan, *et al.*, "Dense hybrid recurrent multi-view stereo net with dynamic consistency checking," in *Proc. ECCV 2020* (Glasgow, UK), Aug. 23-28, 2020, pp. .674–689.

[41] X. Gu, *et al.*, "Cascade cost volume for high-resolution multi-view stereo and stereo matching," in *Proc. CVPR 2020*, Jun. 14-19, 2020, pp. 2495–2504.

[42] J. Yang, W. Mao, J. M. Alvarez, and M. Liu, "Cost volume pyramid based depth inference for multi-view stereo," in *Proc. CVPR 2020*, Jun. 14-19, 2020, pp. 4877–4886.

[43] F. Wang, *et al.*, "PatchmatchNet: Learned multi-view patchmatch stereo," in *Proc. CVPR 2021*, Jun. 19-25, 2021, pp. 14194–14203.

[44] Z. Mi, D. Chang, and D. Xu, "Generalized binary search network for highly-efficient multi-view stereo," in *Proc. CVPR 2022* (New Orleans, LA, USA), Jun. 19-24, 2022, pp. 12991–13000.

[45] H. Xu, *et al.*, "Digging into uncertainty in self-supervised multi-view stereo," in *Proc. ICCV 2021*, Oct. 11-17, 2021, pp. 6078–6087.

[46] J. Yang, J. M. Alvarez, and M. Liu, "Non-parametric depth distribution modelling based depth inference for multi-view stereo," in *Proc. CVPR 2022* (New Orleans, LA, USA), Jun. 19-24, 2022, pp. 8626–8634.

[47] H. Jie, S. Li, S. Gang, and S. Albanie, "Squeeze-and-excitation networks," in *Proc. CVPR 2018* (Salt Lake City, UT, USA), Jun. 18-23, 2018, pp. 7132–7141.

[48] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. CVPR 2018* (Salt Lake City, UT, USA), Jun. 18-23, 2018, pp. 7794–7803.

[49] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. ECCV 2018* (Munich, Germany), Sep. 8-14, 2018, pp. 3–19.

[50] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "BAM: Bottleneck attention module," arXiv:1807.06514, Jul. 2018, pp. 1-14.

[51] R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," in *Proc. ECCV 1994* (Stockholm, Sweden), May 2-6, 1994, pp. 151–158.

[52] T. Y. Lin, *et al.*, "Feature pyramid networks for object detection," in *Proc. CVPR 2017* (Honolulu, HI, USA), Jul. 21-22, 2017, pp. 2117–2125.

[53] Q. Xu and W. Tao, "PVSNet: Pixelwise visibility-aware multi-view stereo network," arXiv:2007.07714, Jul. 2020, pp. 1–11.

[54] B. Ummenhofer, *et al.*, "DeMoN: Depth and motion network for learning monocular stereo," in *Proc. CVPR 2017* (Honolulu, HI, USA), Jul. 21-26, 2017, pp. 5038–5047.

[55] J. R. Chang and Y. S. Chen, "Pyramid stereo matching network," in *Proc. CVPR 2018* (Salt Lake City, UT, USA), Jun. 18-22, 2018, pp. 5410–5418 .

[56] M. Ji, *et al.*, "SurfaceNet: An end-to-end 3D neural network for multi-view stereopsis," in *Proc. CVPR 2017* (Honolulu, HI, USA), Jul. 21-26, 2017, pp. 2307–2315.

[57] X. Ma, *et al.*, "EPP-MVSNet: Epipolar-assembling based depth prediction for multi-view stereo," in *Proc. ICCV 2021*, Oct. 11-17, 2021, pp. 5732–5740.

[58] F. Wang, S. Galliani, C. Vogel, and M. Pollefeys, "IterMVS: Iterative probability estimation for efficient multi-view stereo," in *Proc. CVPR 2022* (New Orleans, LA, USA), Jun. 18-24, 2022, pp. 8606–8615.

[59] Q. Xu and W. Tao, "Multi-scale geometric consistency guided multi-view stereo," in *Proc.CVPR 2019* (Long Beach, CA, USA), Jun. 16-20, 2019, pp. 5483–5492.

[60] J. Zhang, *et al.*, "Visibility-aware multi-view stereo network," in *Proc. BMVC 2020*, Sep. 7-10, 2020, pp. 1–12.

[61] https://www.tanksandtemples.org/details/6813/: using the online reconstruction results evaluation service 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.