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

Yimei Liu, Qing Cai, *Member, IEEE*, Congcong Wang, Jian Yang, Hao Fan, Junyu Dong, *Member, IEEE*, and Sheng Chen, *Life 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], [3], [4], photogrammetry [5], [6] and cartography [7], [8], [9], [10], [11]. Due to the successes of image matching [22], [23], [24] and SFM [14] algorithms that camera pose can be well estimated, the 3D reconstruction from multiple images can be viewed as a problem of dense matching across images. A commonly used paradigm for MVS is the depth map fusion-based approach. These approaches involve estimating a dense depth map for each input image by incorporating multiple co-visible images and then merging these multi-view depth maps to generate the dense reconstruction. Different from single-image depth estimation, MVS estimates per-pixel depth by considering matching costs across a set of discrete depth candidates for each image. By searching for geometrically consistent matches across input views, reliable depths can be obtained, enabling high-quality 3D reconstruction. So far, numerous MVS methods have been proposed using this paradigm, ranging from early traditional methods [12], [13], [14] to recent deep learning-based methods [15], [16], [17], [18], [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], [16], [17], [22], [23], [24], [25] employ Convolutional Neural Networks (CNNs) to incorporate semantic information for more reliable matching, leading to improved performance on various MVS benchmarks [26], [27], [28], [29]. Some approaches, such as the geometry-based methods [8], [10], [11], [19], [30], [31], [32] and the attention-based methods [17], [34], [35], [36], [37], introduce stable geometric clues and finely designed attention mechanisms to alleviate ill-posed matching issues.

The geometry-based methods proposed to explore stable geometric clues of the scene from the captured images, such as edge [31], channel-wise normal curvature [32], and structural affinity features [30]. Then, the learned geometric features of multiple visual angles are employed for cost

Manuscript received 22 July 2023; revised 26 September 2023 and 23 December 2023; accepted 6 March 2024. Date of publication 18 March 2024; date of current version 12 August 2024. This work was supported in part by the National Science Foundation of China under Grant 42106193 and Grant 41927805, and in part by the National Natural Science Foundation of China (NSFC) under Grant 62102285. This article was recommended by Associate Editor C. Yang. *(Corresponding authors: Junyu Dong; Hao Fan.)*

Yimei Liu, Qing Cai, Jian Yang, Hao Fan, and Junyu Dong are with the Department of Information Science and Technology, Ocean University of China, Qingdao 266100, China (e-mail: liuyimei@stu.ouc.edu.cn; cq@ouc.edu.cn; yangjian7669@stu.ouc.edu.cn; fanhao@ouc.edu.cn; dongjunyu@ouc.edu.cn).

Congcong Wang is with the Department of Computer Science and Engineering, Tianjin University of Technology, Tianjin 300222, China (e-mail: congcong_wang@yeah.net).

Sheng Chen is with the School of Electronics and Computer Science, University of Southampton, SO17 1BJ Southampton, U.K., and also with the Department of Information Science and Technology, Ocean University of China, Qingdao 266100, China (e-mail: sqc@ecs.soton.ac.uk).

Data is available online at https://github.com/yimei910110/GA-MVS.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2024.3376692.

Digital Object Identifier 10.1109/TCSVT.2024.3376692

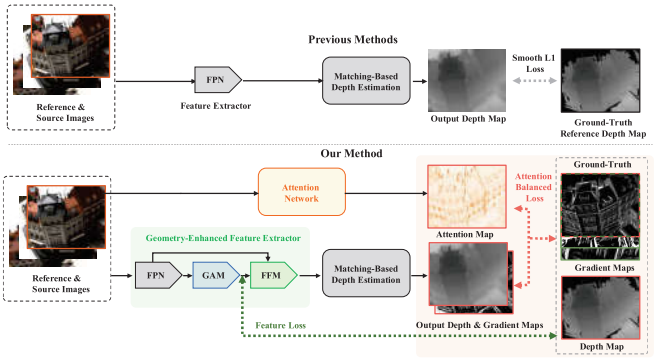


Fig. 1. A brief explanation of our motivation. Different from previous MVS methods, the proposed GA-MVS handles unreliable matching problems via geometry-enhanced feature representation and reliable feature constraint. Collaborative attention network and attention-balanced loss provide adaptive matching measurement for challenging areas, enabling improved matching-based depth estimation performance.

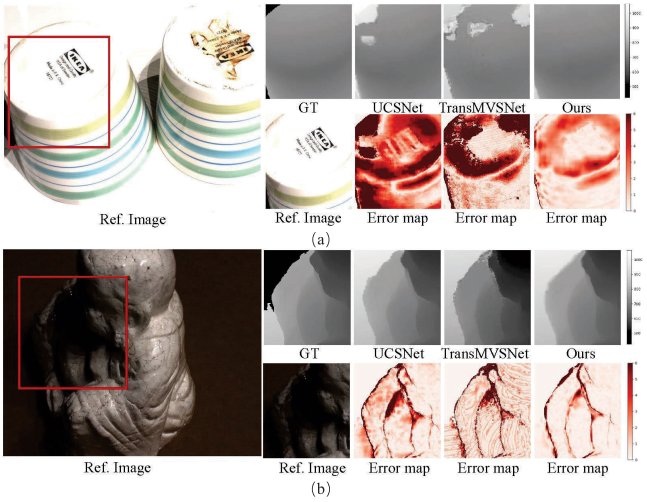


Fig. 2. Performance comparison of learning-based MVS methods on shadow and textureless regions: the baseline model [16], the state-of-the-art transformer-based approach TransMVSNet [34], and our proposed GA-MVS. By incorporating **geometry-enhanced features** and **attentive learning framework**, our method achieves more accurate depth estimates in challenging matching scenarios.

volume construction and depth estimation. The attention-based methods proposed to handle ambiguous matching results via strengthening long-range global context aggregation within and between images [17], [34], [35]. These works achieved promising results but usually introduced high computational complexity. More importantly, these methods only impose constraints on the final regressed or classified depths, indirectly affecting the early-stage extracted features and calculated matching results. This ambiguity implicitly adds the difficulty to robust feature learning in the MVS task. Therefore, the extracted features 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 measurement.

To tackle the aforementioned problems, we propose an advanced framework called Geometry-enhanced Attentive MVS (GA-MVS). We assume that robust MVS reconstruction fundamentally requires reliable feature representations for cross-view correspondence as well as adaptive matching cost evaluations that accommodate real-world ambiguities. We identified these two central factors and proposed corresponding improvements to enhance the robustness of multi-view matching in challenging situations. As shown in Fig. 1, the proposed GA-MVS comprises two crucial improvements: the geometry-enhanced feature extractor and the adaptive matching measurement, achieved by minimizing the attention-balanced loss. For the geometry-enhanced feature extractor, we first extracted one-channel geometric feature from discriminative texture features and introduced accurate depth variance constraint in this step, enabling the extracted geometric features to perceive real geometric clues of the scene, that is, regions with varied depths. Then, we propose a Feature Fusion Module (FFM) to effectively integrate the acquired geometric feature with initial discriminative texture features, thus producing the final geometry-enhanced representations. Our geometry-enhanced features can perceive and weaken pseudo textures caused by illumination and visual angle effects and are consistent across different images. This significantly enhances the depth estimation accuracies. For the adaptive matching cost evaluation, we first estimate a reference depth map and an aligned adaptive attention map using a basic matching-based depth estimation network and a lightweight attention network. Then, pixel-wise adaptive supervisions are learned via minimizing the attention-balanced loss, which is composed of the depth loss term and gradient loss term, weighted by the estimated attention map. Consequently, higher attention weights, which correspond to higher penalties on gradient loss, are applied to textureless regions due to their greater matching uncertainty. In contrast, in well-textured regions, lower gradient losses are applied. Higher penalties on gradient loss influence the depth estimation model to produce depth with less fluctuation. Meanwhile, the depth loss term promotes global correctness. As illustrated in Fig. 2, the baseline and recent transformed-based models both perform biased estimates in challenging areas. In contrast, our model predicts high-qualified depth with fewer biases in both challenging and well-textured regions.

Our main contributions are summarized as follows: -

1) 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.
2) An attentive learning framework is proposed to help the model learn pixel-wise adaptive matching measurement, by minimizing 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.
3) 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. Sec. II provides an overview of related works. Sec. III presents a comprehensive explanation of our proposed methodology, which includes the geometry-enhanced feature extractor and the attentive learning framework. Sec. 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, Sec. 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 Sec. V.

II. Related Work

According to the taxonomy provided in [38], MVS methods can be categorized into four main types: surface evolution-based, voxel-based, point cloud-based, and depth map fusion-based methods. In this section, we will primarily review depth map fusion-based MVS methods, attention mechanisms, and geometric clues utilized in the stereo vision field. These topics are particularly relevant to our technical contributions.

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 ground-breaking depth map-based method, MVSNet [15] proposed a pipeline that includes feature extraction, variance-based feature fusion, and 3D-CNN cost volume regularization steps. This pipeline allows for depth estimation and 3D reconstruction through end-to-end learning procedures. However, there are two main limitations. Firstly, the memory requirement increases significantly with higher input image resolution and depth hypotheses. 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. Secondly, while the cost volume pipeline enables accurate depth estimation in highly textured regions, Lambertian surfaces, and ideal lighting conditions, the depth estimation performance degrades seriously in ill-posed matching situations, such as homogeneous textureless regions, realistic illumination, and view angle effects. 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 [8], 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, we propose incorporating textural and geometric clues to enhance feature representation, and introducing adaptive matching cost evaluation into the conventional cost volume pipeline. Our method provides mutually compatible solutions to concurrently address the critical depth degradation problem in the existing cost volume pipeline.

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], [48], [49], [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. The ambiguous matching results corresponding to low-textured regions are well handled via global context information and inter-image feature interaction, but with high computation complexity. MVSTER [36] and RayMVSNet [18] proposed limiting attention associations within the epipolar line to reduce computation. However, the inconsistent feature representation caused by shadow and reflections still exists. The attention-based models building 3D associations via inter-image feature interaction makes the model vulnerable to erroneous texture similarity caused by multi-view photometric-varied effects.

In our proposed attentive learning framework, unlike previous works that applied attention mechanisms for cost volume construction or regularization, we generate the spatial attention map from the reference image and detected edge features, for the edges are usually associated with textureless regions, oppositely. Then, the appropriate penalty strategies for different areas with varied matching difficulties are learned via the proposed attention-balanced loss, without introducing sophisticated networks, which benefits algorithm efficiency. Besides, multi-view photometric-varied effects are handled by the geometry-enhanced feature representation.

C. Geometric Clues for Learning-Based MVS

Although convolutional features are commonly employed to describe the points and construct matching costs in current

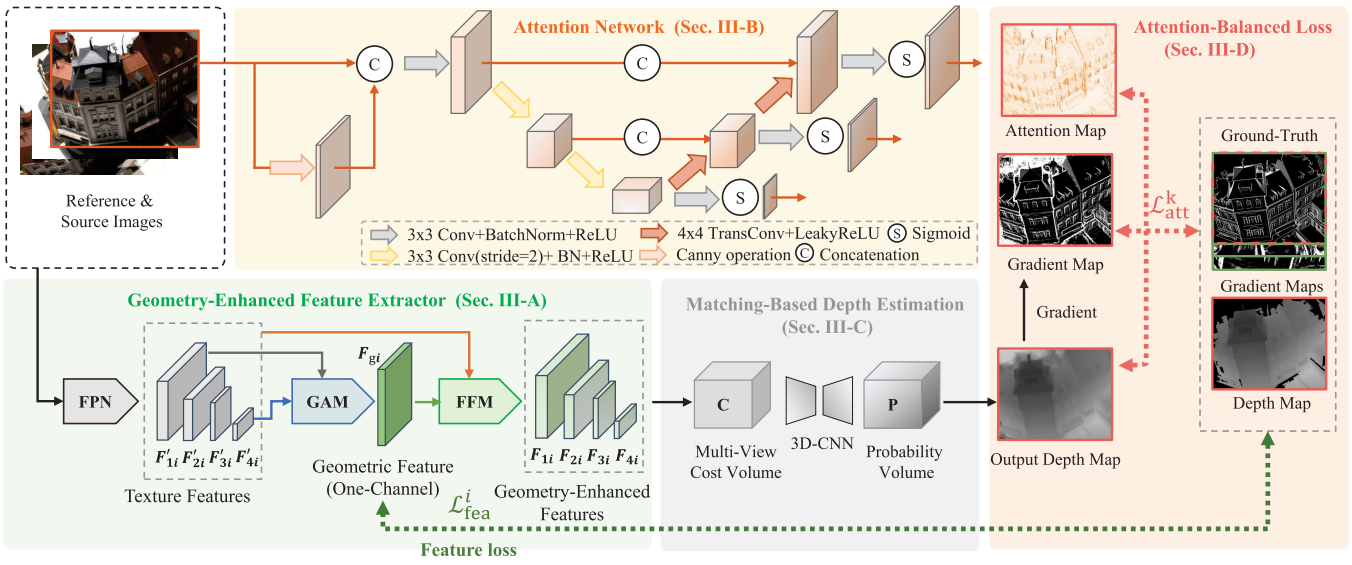


Fig. 3. The overall architecture of GA-MVS. The main components include: (a) Geometry-enhanced feature extractor, which outputs feature maps of input images hierarchically at multiple resolutions, allowing us to advance the depth estimation in a coarse-to-fine manner. (b) Matching-based depth estimation via 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.

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], [31], [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 sub-network 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.

III. Methodology

The overall architecture of our GA-MVS is illustrated in Fig. 3. Firstly, the geometry-enhanced feature extractor is used to obtain multi-scale deep features for robust representation (Sec. III-A). Then, the attention network and the matching-based depth estimation are employed to predict multi-scale attention maps and depth maps, respectively, from coarse to fine (Sec. III-B and Sec. III-C). Finally, we apply the attention-balanced loss with adaptive punishment in ambiguous areas, and feature loss with reliable constraints for geometric features to train GA-MVS end-to-end (Sec. III-D).

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. 3. Specifically, the Feature Pyramid Network (FPN) [52] is first employed 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 Sec. III-A-a), 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 Sec. III-A-b), 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

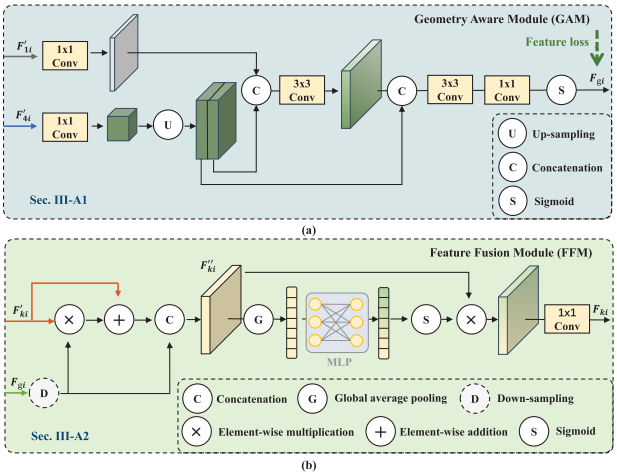


Fig. 4. (a) Illustration of GAM, which explores illumination-invariant geometric feature F_{gi}, where the green dashed arrow indicates the proposed feature loss. (b) Illustration of FFM, which integrates geometric feature F_{gi} into initial texture features F'_{ki} to get stable and discriminative 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.

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 learned implicitly by training end-to-end networks with the proposed feature loss defined in Sec. III-D-a).

The procedure of this GAM is illustrated in Fig. 4 (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 focus on regions with varied depths, which contain real geometric clues. This will become clear in Sec. III-D-a).

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. 4 (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, the critical feature channels is enhanced by performing the effective channel attention mechanism [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:

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

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. 12 of Sec. IV-D-a), 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.

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 Sec. III-D-b). Fig. 3 illustrates the details of 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 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 nonlinearities, 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. Matching-Based Depth Estimation

After the feature extraction step, the cascaded cost volume pipeline [16] is adopted for multi-stage depth estimation 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

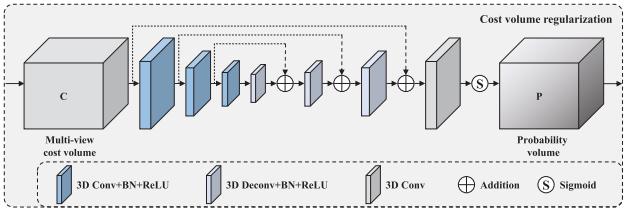


Fig. 5. Network architecture of cost volume regularization step in the matching-based depth estimation. The constructed multi-view cost volume is regularized by a 3D-CNN to get a probability volume for depth regression.

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:

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

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:

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

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, as depicted in Fig. 5, the 3D cost volume C is regularized by a 3D-CNN and transformed into a probability volume $P \in R^{D\times H\times W}$, fomulatted as:

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

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:

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

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:

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

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 $\mathcal{L}^{i}_{fea.}$ 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:

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

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

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

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:

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

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 Sec. IV-D-b)).

The second part of the total loss, $\mathcal{L}_d\big(\widetilde{\mathbf{d}}_{\mathbf{p}}, \mathbf{d}^{gt}_{\mathbf{p}}\big)$, 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:

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

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, with a learning rate of 10^{-4}. 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 Sec. 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], [18], [34], [35], [36], [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. GBINet shows advantages in terms of completeness and overall metrics. We assume that is because their generalized binary search strategy is particularly effective for the DTU small-scale scenes. 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. First, thanks to the direct guidance of feature learning via reliable constraints, our geometry-enhanced features can perceive the real geometric clues of the scene and access 3D-consistent feature representation. Robust feature representation provides the foundation for robust multi-view matching measurement and accurate depth estimation. Second, the proposed attention network and attention-balanced loss help the model recognize unreliable matching cases. Our model has learned the patterns of depth estimation in challenging matching and well-textured regions, with adaptive constraint strategies. Consequently, the estimated depths vary with less fluctuation in unreliable matching areas and lean towards matching results in well-textured regions.

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

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

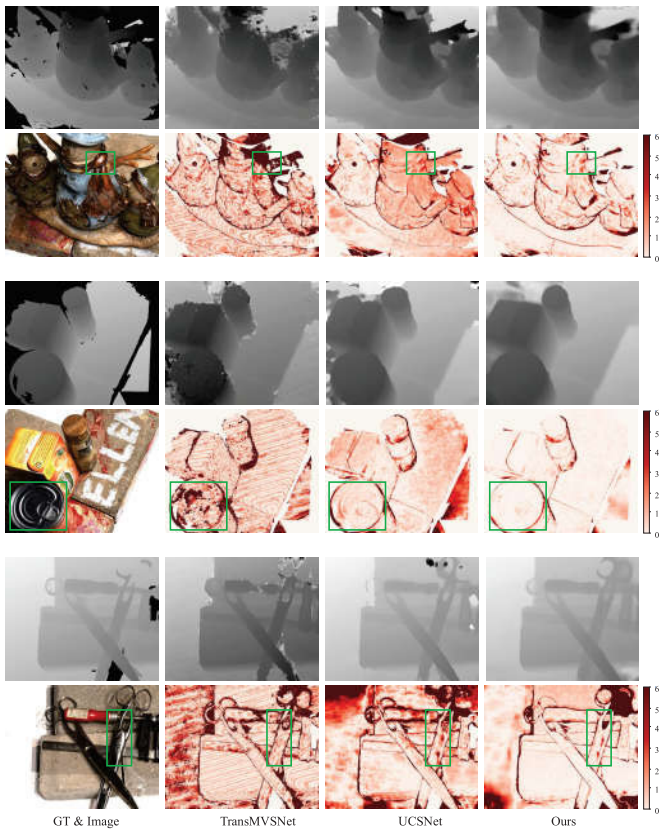


Fig. 6. Visual comparison of estimated depth maps utilizing GA-MVS, TransMVSNet [34] and baseline UCSNet [16]. The green boxes in the observed images indicate regions with specular reflection.

subsets focusing on scenes with rich-textured, textureless, and photometric-varied situations, from the DTU testing set. The rich-textured 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

[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 II
QUANTITATIVE PERFORMANCE ON SCENES WITH DIFFERENT MATCHING DIFFICULTY

	Rich-textured Set	Textureless Set	Photometric-varied Set
UCSNet [16]	0.326	0.371	0.328
GBINet [44]	0.271	0.318	0.304
Ours	0.311	0.313	0.301
Improvement Ratio	4.4%	15.8%	8.4%

The evaluation metric for the first three rows is the averaged error of reconstructed point clouds, measured in millimeters. The Improvement Ratio row indicates the gain compared to the baseline model UCSNet.

TABLE III
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.

shadow. We compare the reconstruction results of our model, the baseline UCSNet [16], and the recent advanced model GBINet [44] in Table II. When compared to UCSNet, the results show that the proposed method is capable of alleviating the depth degradation problem in the existing cost volume pipeline, achieving robust multi-view matching and enhanced reconstruction performance in varied matching difficulty, especially for challenging textureless and photometric-varied cases. The improvement ratios are 15.8% and 8.4% compared to UCSNet. When compared to GBINet, our model performs worse on the rich-textured set but outperforms GBINet on the more challenging textureless and photometric-varied sets. This confirms that our model can achieve high-qualified reconstruction in challenging matching situations. Meanwhile, GBINet shows advantages in reconstructing general highly textured scenes via their generalized binary search strategy. We visualize some reconstructed results on the DTU testing set [26] by our method in Fig. 7. Fig. 8 provides corresponding qualitative error results. Our reconstructions are dense and accurate in the details of the scene.

Table III 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.

2) Generalization on Tanks & Temples Benchmark: To verify the generalization capability of our method, First, we fine-tuned the pre-trained model on DTU by using BlendedMVS dataset. Second, we used directly the model trained on BlendedMVS datasets for evaluation, followed the experimental settings of UGNet [8]. 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 14 learning-based MVS methods. We observe that the proposed GA-MVS obtains competitive F-scores on both sets, indicating the strong generalization of our model. Fig. 9 compares the respective error

Fig. 7. Reconstruction results on DTU's testing set by our proposed approach.

Fig. 8. Qualitative error results correspond to Fig. 7. The variation from white to red in increasing order illustrates the magnitude of the error. The points marked blue/green are masked out in the evaluation, for lacking corresponding scan data as ground-truth.

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
GBINet◇ [44]	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
UGNet† [8]	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
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
GA-MVS (Ours)†	2023	**63.95**	80.57	67.06	**57.70**	66.15	62.25	60.58	60.11	57.19	**38.94**	25.14	**46.06**	37.39	50.91	34.05	**40.11**

* indicates traditional MVS methods, the others are learning-based MVS methods. ○ indicates only training on the training set of DTU. † indicates training on BlendedMVS. Others are trained on DTU and then fine-tuned on BlendedMVS. 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].

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. 10 illustrates the reconstructed point clouds on the Tanks & Temples benchmark achieved by our method. Fig. 11 provides corresponding error visualization. Our reconstruction results are complete and with rich details in various scenes, demonstrating its generalization capability.

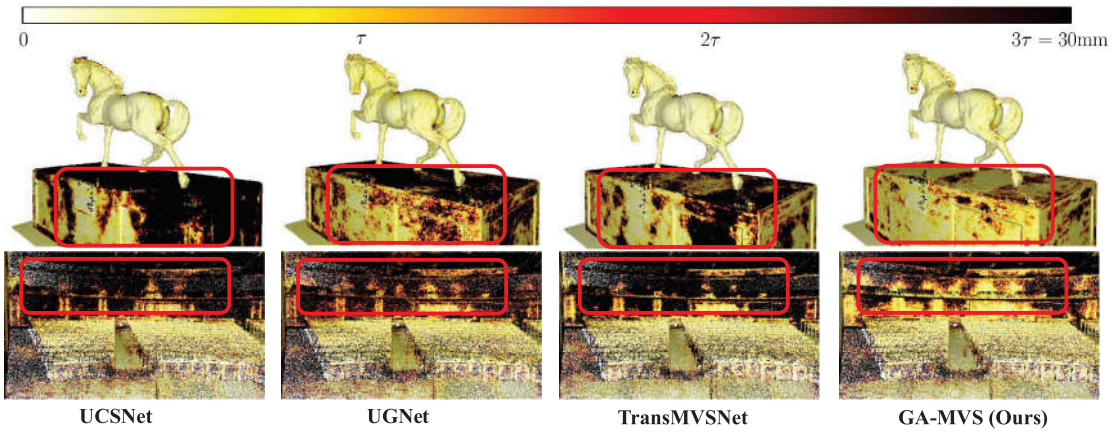


Fig. 9. Error visualization of Horse and Auditorium scenes in the Tanks & Temples benchmark [27]. We exemplify the errors of baseline UCSNet [16], two best performing existing models UGNet [8] and TransMVSNet [34], and ours, computed based on ground-truth point clouds. Darker points indicate bigger errors of reconstructions.

Fig. 10. Reconstruction results on Tanks & Temples dataset by our proposed approach.

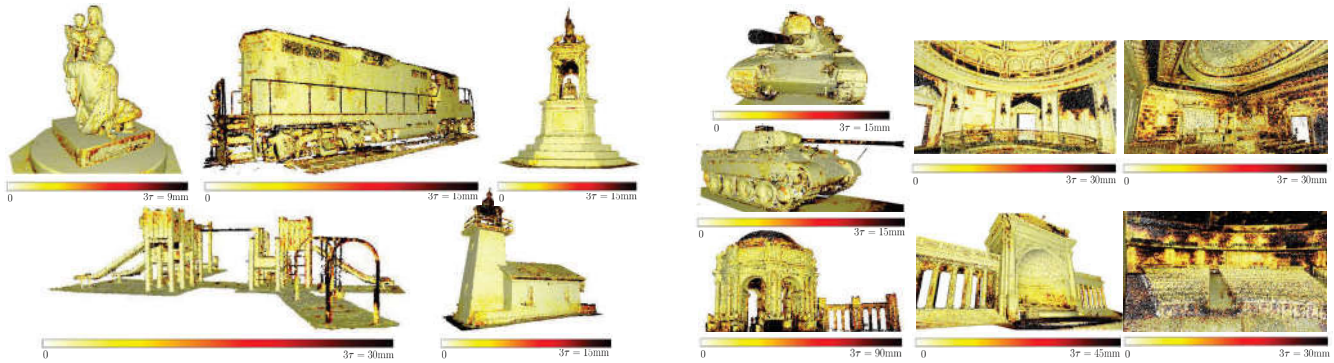


Fig. 11. Error visualization corresponds to Fig 10.

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. The ablated results are shown in Table VI. We witness that both cascade cost volume baselines demonstrate higher 3D reconstruction performance in terms of point cloud accuracy and completeness quality. This is achieved through the proposed geometry-enhanced feature extractor and adaptive matching measurement. These findings highlight the practical contributions in addressing the depth estimation degradation problem in the existing cost

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$\times$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 Sec. III-A, and Att. Loss indicates attention-balanced loss detailed in Sec. III-D. R. column indicates the improvement ratio, in terms of overall quality, by adding one more component.

volume pipeline. Then, we conduct ablation experiments on two contributions to get more insights into how they play parts step-by-step.

1) Geometry-Enhanced Features: For the performance improvement brought by the geometry-enhanced features, we consider that there are two potential 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.

a) 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 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 manner 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.

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

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

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 $\frac{W}{2^{k-1}} \times \frac{H}{2^{k-1}}$

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.

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. To keep other modules unchanged, for experiments of $F'_{2i}+F'_{3i}$, $F'_{2i}+F'_{4i}$, $F'_{3i}+F'_{4i}$ as the input, the corresponding output of GAM is adjusted to be the same as the input image resolution via an extra upsample operation. 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 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 reconstruction accuracy 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. We infer this benefits from the initial fusion step of the following FFM. The initial fusion step of FFM enhances channels containing varied depths while the others remain unchanged. Therefore, although the output geometric feature contains imperfections, the input discriminative texture features still remain to ensure qualified matching results.

b) Visualization of geometry-enhanced features: We compare the geometry-enhanced features with the FPN-extracted texture features in Fig. 12. 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.

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

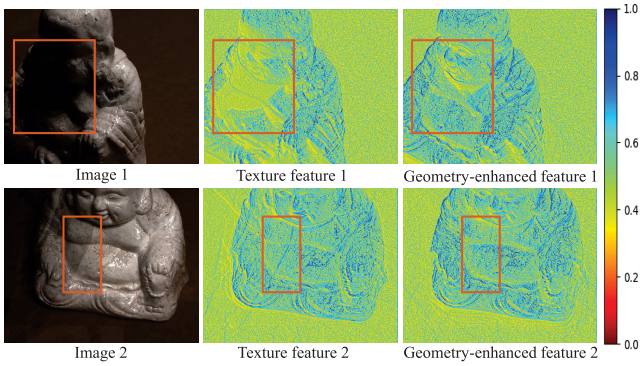


Fig. 12. Comparison of initial texture features and geometry-enhanced features corresponding to two input images. Our geometry-enhanced features can effectively restrain view-dependent photometric effects in shadow and shadow boundary regions by reasonably introducing reliable depth variance constraints in the early feature extraction step.

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.

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 matching-based 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.

a) 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,

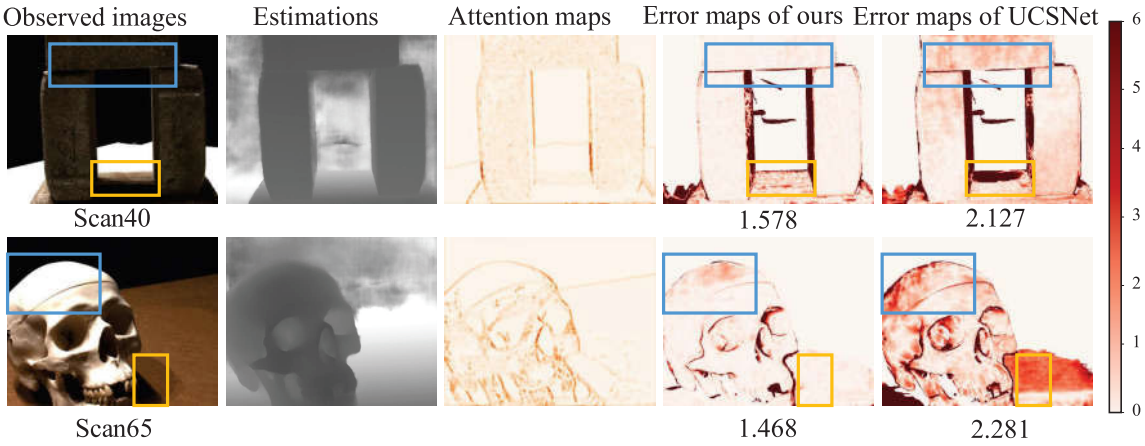


Fig. 13. Visual comparisons for images of scan40 and scan65 scenes. Blue boxes indicate textureless regions, while yellow boxes show shadow regions. Numbers under depth error maps indicate their MAE.

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.

b) Visualization of attention-balanced loss: Fig. 13 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. 13 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, the baseline model cannot distinguish between reliable and unreliable matching results, leading to large errors in challenging textureless regions.

c) Ablation study of loss weight λ: We carry out an ablation study on the hyperparameter λ in Eq. (10). As illustrated in Fig. 14, $\lambda = 8$ is an appropriate weight for the proposed attention-balanced loss.

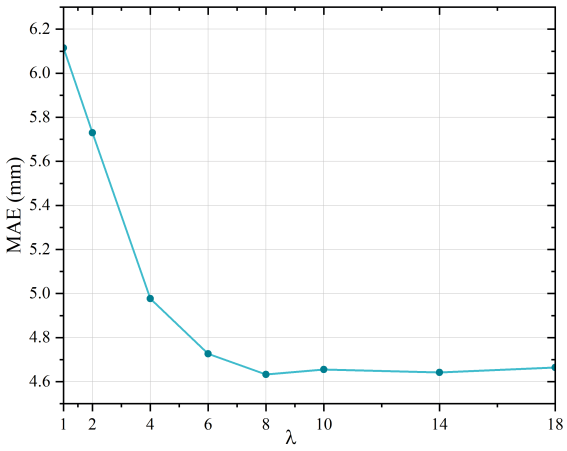


Fig. 14. Ablation study on hyperparameter λ.

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. Compared with the baseline setting, we decrease the number of depth candidates for partial stages and increase the number of stages to get a trade-off for the achievable performance with computational complexity.

a) 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 the setting of depth candidate and feature channel numbers listed in the table. 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.

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.

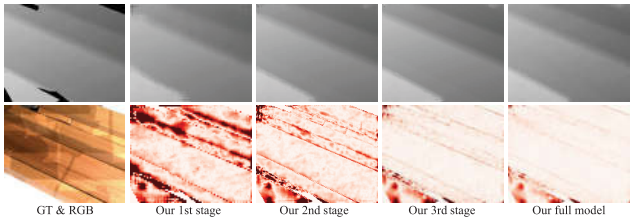


Fig. 15. Depth inference results of each stage. Top-row: Ground truth depth map and intermediate depth inference results. Bottom row: Reference image and error maps of intermediate results.

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.

b) Achievable performance at different stages: To investigate our model's achievable performance at different stages, 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. 15. The overall quality is enhanced from 0.762 to 0.309 in a coarse-to-fine manner. Accordingly, the GPU memory increases from 1161 MB to 6791 MB, and the run-time increases from 0.088 s to 0.757 s.

V. CONCLUSION

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. Park, and H. 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. Syst., Man, Cybern., Syst.*, 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. Cybern.*, vol. 53, no. 3, pp. 1682–1698, Mar. 2023.
[5] F. Rottensteiner, G. Sohn, M. Gerke, J. D. Wegner, U. Breitkopf, and J. Jung, "Results of the ISPRS benchmark on urban object detection and 3D building reconstruction," *ISPRS J. Photogramm. Remote Sens.*, vol. 93, pp. 256–271, Jul. 2014.
[6] S. Malihi, M. J. Valadan Zoej, M. Hahn, M. Mokhtarzade, and H. Arefi, "3D building reconstruction using dense photogrammetric point cloud," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. B3, pp. 71–74, Jun. 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. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 2, pp. 97–102, May 2018.
[8] W. Su, Q. Xu, and W. Tao, "Uncertainty guided multi-view stereo network for depth estimation," *IEEE Trans. Circuits Syst. Video Technol.*, 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 Sens.*, vol. 12, no. 13, p. 2128, Jul. 2020.
[10] H. Dai, X. Zhang, Y. Zhao, H. Sun, and N. Zheng, "Adaptive disparity candidates prediction network for efficient real-time stereo matching," *IEEE Trans. Circuits Syst. Video Technol.*, 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 Syst. Video Technol.*, early access, May 16, 2023, doi: 10.1109/TCSVT.2023.3276753.
[12] C. Bailer, M. Finckh, and H. Lensch, "Scale robust multi view stereo," in *Proc. ECCV*, Florence, Italy, 2012, pp. 398–411.
[13] (2015). *Open Multi-View Stereo Reconstruction Library*. [Online]. Available: https://github.com/cdcseacave/openMVS
[14] J. L. Schönberger, E. Zheng, J. M. Frahm, and M. Pollefeys, "Pixelwise view selection for unstructured multi-view stereo," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, 2016, pp. 501–518.
[15] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "MVSNet: Depth inference for unstructured multi-view stereo," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, 2018, pp. 767–783.
[16] S. Cheng et al., "Deep stereo using adaptive thin volume representation with uncertainty awareness," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2524–2534.
[17] K. Luo, T. Guan, L. Ju, Y. Wang, Z. Chen, and Y. Luo, "Attention-aware multi-view stereo," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1587–1596.
[18] J. Xi, Y. Shi, Y. Wang, Y. Guo, and K. Xu, "RayMVSNet: Learning ray-based 1D implicit fields for accurate multi-view stereo," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 8595–8605.
[19] 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 Trans. Circuits Syst. Video Technol.*, early access, Apr. 17, 2023, doi: 10.1109/TCSVT.2023.3267457.

[20] H. Liu, Q. Zhang, B. Fan, Z. Wang, and J. Han, "Features combined binary descriptor based on voted ring-sampling pattern," *IEEE Trans. Circuits Syst. Video Technol.*, 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. Syst., Man, Cybern., Syst.*, 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 Process.*, vol. 31, pp. 4842–4855, 2022.

[23] B. Fan, Y. Yang, W. Feng, F. Wu, J. Lu, and H. Liu, "Seeing through darkness: Visual localization at night via weakly supervised learning of domain invariant features," *IEEE Trans. Multimedia*, vol. 25, pp. 1713–1726, 2023.

[24] B. Fan, H. Liu, H. Zeng, J. Zhang, X. Liu, and J. Han, "Deep unsupervised binary descriptor learning through locality consistency and self distinctiveness," *IEEE Trans. Multimedia*, vol. 23, pp. 2770–2781, 2021.

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

[26] R. Jensen, A. Dahl, G. Vogiatzis, E. Tola, and H. Aanæs, "Large scale multi-view stereopsis evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 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. Graph.*, vol. 36, no. 4, pp. 1–13, 2017.

[28] T. Schops et al., "A multi-view stereo benchmark with high-resolution images and multi-camera videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 3260–3269.

[29] Y. Yao et al., "BlendedMVS: A large-scale dataset for generalized multi-view stereo networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 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, 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. Asian Conf. Comput. Vis.*, 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, pp. 1–19.

[33] Z. Wei, Q. Zhu, C. Min, Y. Chen, and G. Wang, "AA-RMVSNet: Adaptive aggregation recurrent multi-view stereo network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 6187–6196.

[34] Y. Ding et al., "TransMVSNet: Global context-aware multi-view stereo network with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 8585–8594.

[35] A. Yu et al., "Attention aware cost volume pyramid based multi-view stereo network for 3D reconstruction," *ISPRS J. Photogramm. Remote Sens.*, vol. 175, pp. 448–460, May 2021.

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

[37] J. Zhu, B. Peng, W. Li, H. Shen, Z. Zhang, and J. Lei, "Multi-view stereo with transformer," 2021, *arXiv:2112.00336*.

[38] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, New York, NY, USA, Jun. 2006, pp. 1–8.

[39] Y. Yao, Z. Luo, S. Li, T. Shen, T. Fang, and L. Quan, "Recurrent MVSNet for high-resolution multi-view stereo depth inference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5525–5534.

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

[41] X. Gu, Z. Fan, S. Zhu, Z. Dai, F. Tan, and P. Tan, "Cascade cost volume for high-resolution multi-view stereo and stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 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. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4877–4886.

[43] F. Wang, S. Galliani, C. Vogel, P. Speciale, and M. Pollefeys, "PatchmatchNet: Learned multi-view patchmatch stereo," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14194–14203.

[44] Z. Mi, C. Di, and D. Xu, "Generalized binary search network for highly-efficient multi-view stereo," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 12981–12990.

[45] H. Xu et al., "Digging into uncertainty in self-supervised multi-view stereo," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6058–6067.

[46] J. Yang, J. M. Alvarez, and M. Liu, "Non-parametric depth distribution modelling based depth inference for multi-view stereo," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 8616–8624.

[47] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[48] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[49] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, Sep. 2018, pp. 3–19.

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

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

[52] T. Y. Lin, P. Dollàr, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2117–2125.

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

[54] B. Ummenhofer et al., "DeMoN: Depth and motion network for learning monocular stereo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 5622–5631.

[55] J. R. Chang and Y. S. Chen, "Pyramid stereo matching network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 5410–5418.

[56] M. Ji, J. Gall, H. Zheng, Y. Liu, and L. Fang, "SurfaceNet: An end-to-end 3D neural network for multiview stereopsis," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Honolulu, HI, USA, Oct. 2017, pp. 2307–2315.

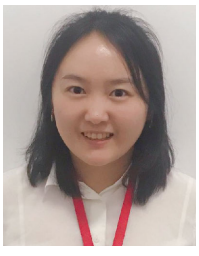
[57] X. Ma, Y. Gong, Q. Wang, J. Huang, L. Chen, and F. Yu, "EPP-MVSNet: Epipolar-assembling based depth prediction for multi-view stereo," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 5732–5740.

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

[59] Q. Xu and W. Tao, "Multi-scale geometric consistency guided multi-view stereo," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 5483–5492.

[60] J. Zhang, Y. Yao, S. Li, Z. Luo, and T. Fang, "Visibility-aware multi-view stereo network," in *Proc. Brit. Mach. Vis. Conf.*, 2020, pp. 1–12.

[61] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun. (2017). *Tanks and Temples: Benchmarking Large-scale Scene Reconstruction*. [Online]. Available: https://www.tanksandtemples.org/details/6813/



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

Qing Cai (Member, IEEE) received the M.Sc. and Ph.D. degrees from the Department of Automation, Northwestern Polytechnical University, Xi'an, China, in 2016 and 2019, respectively. From 2017 to 2018, he was a Visiting Ph.D. Student with the Department of Computing Science, University of Alberta. From 2020 to 2022, he was a Post-Doctoral Fellow with The Chinese University of Hong Kong Shenzhen and the University of Science and Technology of China. He is currently an Associate Professor with the Faculty of Information Science and Engineering, Ocean University of China. His research interests include machine learning, deep learning, and computer vision, with a focus on image restoration, image segmentation, medical image processing, and visual tracking.

Congcong Wang received the M.Sc. degree in Erasmus mundus master program color in informatics and media technology (ClMET) in 2014 and the Ph.D. degree in computer science from the Norwegian University of Science and Technology, Norway, in 2020. She is currently a Lecturer with Tianjin University of Technology.

Jian Yang received the B.Sc. degree in science information and the M.Sc. degree in electronic science and technology in 2015 and 2018, respectively. She is currently pursuing the Ph.D. degree with the Ocean University of China. Her research interests include computer vision, visual localization, and machine learning.

Hao Fan received the B.Sc., M.E., and Ph.D. degrees from the Department of Computer Science and Technology, Ocean University of China, Qingdao, China, in 2012, 2014, and 2019, respectively. He is currently a Lecturer 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 (Member, IEEE) 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 (Life Fellow, IEEE) received the B.Eng. degree in control engineering from the East China Petroleum Institute, Dongying, China, in 1982, the Ph.D. degree in control engineering from the City University of London, London, in 1986, and the Doctor of Sciences (D.Sc.) degree from the University of Southampton, Southampton, U.K., in 2005. From 1986 to 1999, he held research and academic appointments with the Universities of Sheffield, Edinburgh, and Portsmouth, U.K. Since 1999, he has been with the School of Electronics and Computer Science, University of Southampton, where he is currently a Professor in intelligent systems and signal processing. His research interests include machine learning, neural networks, and wireless communications. He has published over 700 research papers. He has more than 19,000 Web of Science citations with H-index 61 and more than 38,000 Google Scholar citations with H-index 83. He 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 in March 2004.