2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2025)
October 19-25, 2025. Hangzhou, China

ACP-MVS: Efficient Multi-View Stereo with Attention-based Context Perception

Hao Jia[1], Gangwei Xu[1], Miaojie Feng[1], Xianqi Wang[2], JunDa Cheng[1], Min Lin[2] and Xin Yang[1*]

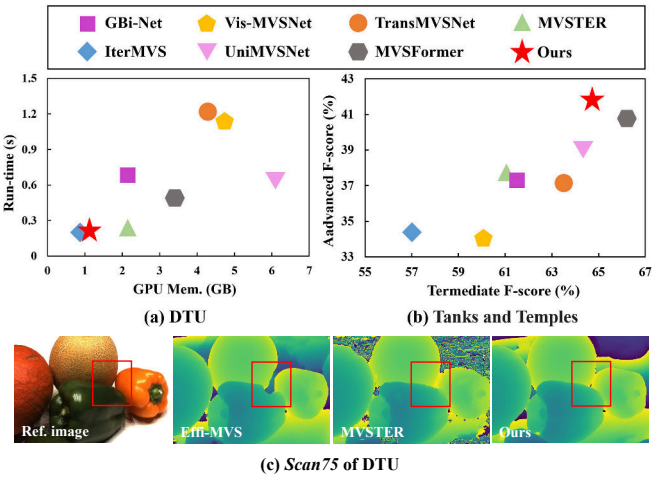
***Abstract*— The core of Multi-View Stereo (MVS) is to find corresponding pixels in neighboring images. However, due to challenging regions in input images such as untextured areas, repetitive patterns, or reflective surfaces, existing methods struggle to find precise pixel correspondence therein, resulting in inferior reconstruction quality. In this paper, we present an efficient context-perception MVS network, termed ACP-MVS. The ACP-MVS constructs a context-aware cost volume that can enhance pixels containing essential context information while suppressing irrelevant or noisy information via our proposed Context-stimulated Weighting Fusion module. Furthermore, we introduce a new Context-Guided Global Aggregation module, based on the insight that similar-looking pixels tend to have similar depths, which exploits global contextual cues to implicitly guide depth detail propagation from high-confidence regions to low-confidence ones. These two modules work in synergy to substantially improve reconstruction quality of ACP-MVS without incurring significant additional computational and time cost. Extensive experiments demonstrate that our approach not only achieves state-of-the-art performance but also offers the fastest inference speed and minimal GPU memory usage, providing practical value for practitioners working with high-resolution MVS image sets. Notably, our method ranks 2nd on the challenging Tanks and Temples advanced benchmark among all published methods. Code is available at https://github.com/HaoJia-mongh/ACP-MVS.**

I. INTRODUCTION

Multi-View Stereo (MVS) is one of the core branches of three-dimensional (3D) computer vision, which aims to reconstruct the 3D geometry of a scene from a collection of 2D overlapping images with known camera parameters. Over the past few years, this task has been extensively studied due to its widespread applications in areas such as autonomous driving, robot navigation and augmented reality. Traditional MVS methods [1], [2], [3] have delivered exceptional results in terms of reconstruction quality. Recently, with advances in deep learning, learning-based MVS methods have emerged and demonstrated superior performance and robustness, especially when dealing with challenging regions such as textureless areas or reflective surfaces. These methods have set new standards and are gradually evolving into the dominant force in the field.

[1]Hao Jia, Gangwei Xu, Miaojie Feng, JunDa Cheng and Xin Yang* are with the Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, 430074, China. (* represents the corresponding author. E-mail: haojia@hust.edu.cn; gwxu@hust.edu.cn; fmj@hust.edu.cn; jundacheng@hust.edu.cn; xinyang2014@hust.edu.cn)

[2]Xianqi Wang and Min Lin are with Institute of Artificial Intelligence, Huazhong University of Science and Technology, Wuhan, 430074, China. (E-mail: xianqiw@hust.edu.cn; minlin@hust.edu.cn)



(c) *Scan75* of DTU

Fig. 1. (a) Run-time vs GPU Mem. on DTU [4] (1600×1152 images, 5 views). (b) Quantitative comparison on Tanks and Temples benchmark [5] (higher is better). (c) Qualitative comparison of estimated depth maps with state-of-the-art efficient methods [6], [7]. ACP-MVS handles challenging regions well, recovering complete depth maps with distinct edges.

The essence of MVS is addressing a densely pixel-wise matching problem which entails locating the corresponding pixel in the reference image along the epipolar line in all neighboring images. Learning-based methods [8], [9] typically extract image features from input images and construct a cost volume encoding matching costs across different depth hypotheses via the plane sweep algorithm [10]. Subsequently, this volume is subject to regularization to produce the final depth estimation. However, challenging regions such as untextured areas, repetitive patterns, or inconsistent illumination, make it difficult to correctly identify the corresponding pixels between the reference image and all source images.

Currently, several learning-based MVS methods have made some progress in solving the aforementioned problem. [11] introduce deformable convolutions to expand the receptive field, and [12] leverage extra CNN networks to learn per-pixel weights as fusing guidance. However, limited by CNN's local receptive fields, these methods still have difficulties in handling large-scale repetitive patterns or textureless regions. Inspired by Transformers' ability to model long-range dependencies, TransMVSNet [13] and MVSFormer [14] introduce feature matching transformers to extract dense features aggregating long-range context information. Additionally, WT-MVSNet [15] and CostFormer [16] propose innovative Transformer-based cost aggregation to refine depth estimation by replacing 3D CNN networks. However, using stacked attention modules significantly increases memory usage and inference time, which poses

challenges for networks processing high-resolution image datasets on mainstream GPU devices. Consequently, a critical question arises: How might we efficiently utilize context information to alleviate pixel-wise mismatching problems without incurring substantial additional costs?

In this paper, we present an efficient MVS network with attention-based context perception, named ACP-MVS, which leverages a Context-Stimulated Weighting Fusion (CWF) module to construct a context-perception cost volume. CWF assigns larger weights to pixels with critical context information and smaller weights to those likely to cause matching ambiguity. Specifically, the CWF module first processes the two-view cost volume through a lightweight 3D CNN network to generate adaptive per-view initial weights and then stimulates these weights using attention maps derived from reference-image context features. Per-view weights encode similarity information, while context features provide geometric and texture cues, enabling the weights to better capture the reference image's geometry information. Attention maps are shared across depth hypotheses, ensuring comprehensive geometric understanding while maintaining computational efficiency. These context-stimulated weights not only capture the varying importance within each view-wise volume and between different view-wise volumes, but also incorporate pixel similarities across all depth hypotheses and geometric cues from the reference-image context feature. As a result, they can effectively guide the fusion of two-view volumes into a multi-view cost volume.

In addition, we propose a Context-Guided Global Aggregation (CGA) module. Based on the insight that pixels with similar appearances tend to have similar depths, this module implicitly propagates depth information from high-confidence to low-confidence regions through global context cues. Specifically, this module employs linear attention [17] to compute attention maps based on the self-similarities of reference-image context features, to guide and enhance volume aggregation with global awareness. Due to its lightweight nature, linear attention enables efficient processing of high-resolution images at a negligible cost. CWF and CGA cooperate synergistically: CWF constructs a context-stimulated cost volume to enhance matching accuracy. Following this, CGA propagates this enhanced information using attention-based context, effectively addressing matching ambiguity. This is validated by experimental results. These modules significantly improve reconstruction quality with minimal computational and time overhead.

Our method achieves state-of-the-art performance in terms of accuracy, inference speed, and memory efficiency, as illustrated in Figure 1. Compared to MVSFormer [14], our method achieves a remarkable 66.9% reduction in GPU memory usage and an impressive 59.2% reduction in inference time, while maintaining comparable reconstruction quality. Among all published methods, ACP-MVS ranks second on the advanced sequences of the Tanks and Temples benchmark [5]. Notably, ACP-MVS is the most efficient among the top 10 methods on DTU [4] and Tanks and Temples in terms of speed and GPU memory. Our networks are simple, accurate, and efficient, offering valuable utility for practitioners working with high-resolution MVS image sets.

II. Related Work

A. Learning-based Multi-View Stereo

Driven by the success of deep learning, MVS has achieved significant progress over traditional methods. Yao et al. [8] pioneer a widely recognized end-to-end MVS pipeline that constructs the cost volume by encoding deep features and camera parameters. They further regularize the cost volume using 3D CNNs to infer depth maps. To reduce the computational cost of the 3D U-Net architecture, subsequent studies have proposed various methods, including coarse-to-fine depth optimization methods [9], [18], recurrent methods [19], [11] based on RNNs, and patch-matching-based methods [20], depending on different regularization patterns. Despite their success, these methods encounter difficulties in challenges such as low-texture and untextured regions.

B. Attention-based Multi-View Stereo

Attention mechanism [21], which was initially designed for natural language processing, has been extensively explored in the visual community as well [22]. Due to its natural superiority to capture long-range dependencies, attention has been widely adopted in MVS. EPP-MVSNet [23] employs epipolar transformers to model spatial relationships. TransMVSNet [13] and MVSFormer [14] introduce feature matching transformers and pre-trained vision transformers, respectively, achieving remarkable results. CostFormer [16] proposes Transformer-based cost aggregation methods to improve accuracy. However, attention-based methods are computationally expensive, limiting their applicability to high-resolution images on mainstream GPU devices.

C. Efficient Multi-View Stereo

To enhance efficiency, several methods adopt a coarse-to-fine strategy, reducing the number of depth hypotheses as resolution increases. Specifically, CVP-MVSNet [9] and CasMVSNet [18] utilize cascading cost volumes. PatchmatchNet [20] introduces the concept of Patchmatch, significantly reducing runtime and memory costs. GBiNet [24] proposes a generalized binary search network to minimize depth hypotheses per stage. Effi-MVS [7] and IterMVS [25] introduce GRU-based iterative architectures for improved efficiency. However, cascade methods struggle to mitigate cumulative errors from coarser resolutions.

III. Method

Given the limited local receptive field of CNN and the memory burden of the Transformer, there has been limited research on effectively and affordably leveraging context information to deal with pixel-wise challenging regions. In this paper, we propose a lightweight attention-based method to adequately utilize reference-image context information, globally guiding cost volume construction and aggregation. Specifically, we introduce the Context-Stimulated Weighting

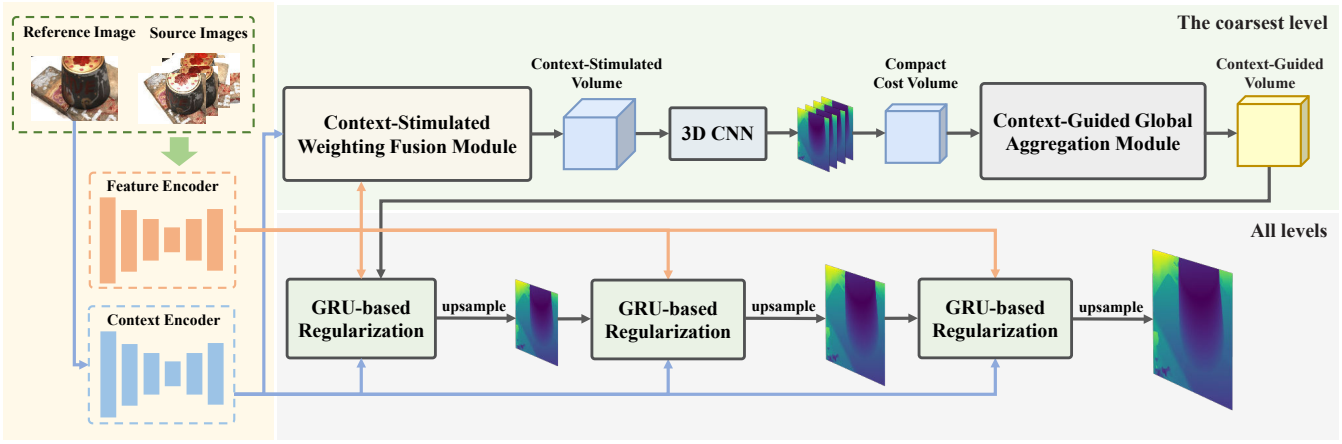


Fig. 2. **Overview of the ACP-MVS.** Our network uses two feature extractors to extract multi-scale image and context features. CWF and CGA integrate context information into the context-perception cost volume construction and aggregation at the coarsest level. A multi-stage GRU-based regularization iteratively updates depth maps.

Fusion (CWF) module (Sec III-A) and the Context-Guided Global Aggregation (CGA) module (Sec III-B). These modules significantly enhance the reconstruction quality of ACP-MVS(Sec III-C).

A. Context-Stimulated Weighting Fusion

Two-view cost volume construction. We construct two-view volumes by warping source-image features into the reference image's camera coordinates based on sampled depth hypotheses. Specifically, for each pixel p in the reference-image feature F_{ref}, we use differentiable homography to warp source-image feature maps F_i and compute $\mathrm{p}_{i,j}$:

$$\mathrm{p}_{i,j} = \mathrm{K}_i \cdot \left(\mathrm{R}_{ref,i} \cdot \left(\mathrm{K}_{ref}^{-1} \cdot \mathrm{p} \cdot \mathrm{d}_j\right) + \mathrm{t}_{ref,i}\right), \quad (1)$$

where K_{ref} and K_i are the intrinsic matrices of F_{ref} and F_i, respectively. $\mathrm{R}_{ref,i}$ and $\mathrm{t}_{ref,i}$ represent the relative rotation and translation matrices, and d_j is the j-th depth hypothesis. Given $\mathrm{p}_{i,j}$ and F_i, we use differential bilinear interpolation [26] to reconstruct the warped feature map F_i'. Subsequently, the warped feature maps for all depth hypotheses are concatenated together as the feature volume $\mathrm{V}_i \in \mathbb{R}^{C\times D\times H\times W}$, where H, W and C represent the height, width and channel dimension of the feature map, respectively, and D is the number of depth hypotheses. Finally, the two-view cost volume $\mathrm{C}_{ref,i}$ is computed as the squared difference between V_i and the reference feature volume V_{ref}, representing the matching information between the two features, as follows:

$$\mathrm{C}_{ref,i} = (\mathrm{V}_{ref} - \mathrm{V}_i)^2. \quad (2)$$

Context-stimulated weighting volume fusion. After constructing two-view cost volumes, the next step is to fuse them into a unified multi-view cost volume for regularization. Conventional methods typically create an indiscriminate multi-view cost volume. However, treating all views equally can make the matching process vulnerable to issues like inconsistent illumination due to varying camera positions. Therefore, we propose the Context-Stimulated Weighting Fusion (CWF) module to learn the importance of volumes from different views, as illustrated in Figure 3. First, a context extractor extracts the context feature $\mathrm{F}_c \in \mathbb{R}^{C\times H\times W}$ from the reference image. Simultaneously, a lightweight 3D CNN processes the two-view cost volume $\mathrm{C}_{ref,i} \in \mathbb{R}^{C\times D\times H\times W}$ to adaptively generate the initial weight $\omega_i \in \mathbb{R}^{1\times D\times H\times W}$. Subsequently, F_c is passed through a lightweight sub-network to produce an attention map $\alpha \in \mathbb{R}^{1\times 1\times H\times W}$ for ω_i. The context-stimulated weight is computed as follows:

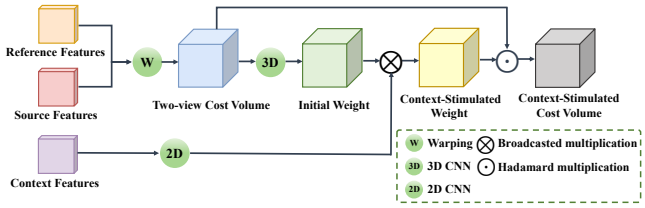


Fig. 3. **Context-Stimulated Weighting Fusion Module.** This module uses attention maps computed from context features to stimulate two-view volume information, serving as context-stimulated weights to guide volume construction. The figure shows the two-view volume construction process.

$$\omega_i = f^{3D}(\mathrm{C}_{ref,i}), \quad (3)$$

$$\alpha = f^{2D}(\mathrm{F}_c), \quad (4)$$

$$\omega_i^c = \alpha \times \omega_i, \quad (5)$$

where $\times$ denotes broadcasted multiplication, f^{3D} and f^{2D} denote 3D and 2D point-wise CNNs, respectively. The attention map is shared along the depth channel. ω_i^c encodes similarity information derived from the cost volume, while the attention map stimulates geometric and textural features by sharing α exclusively along the depth channel, thereby enabling ω_i^c to comprehensively capture geometric cues from the reference image. Consequently, the context-stimulated weight enhance pixels containing critical context information while suppressing low-confidence regions during matching. The final fused volume is defined as follows:

$$\mathrm{C}_{CWF} = \frac{1}{N-1}\sum_{i=1}^{N-1} \omega_i^c \odot \mathrm{C}_{ref,i}, \quad (6)$$

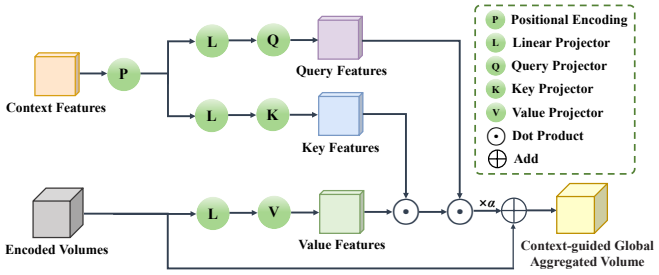


Fig. 4. **Context-Guided Global Aggregation Module.** This module encodes reference-image context features into query and key features, respectively. It computes the dot product between the encoded volume and these two features individually, sums the final results with the encoded volume, and generates the context-guided global aggregated volume.

where $\odot$ denotes Hadamard multiplication. This module is simple to implement, requiring minimal additional operations while significantly improving performance.

B. Context-Guided Global Aggregation

Based on the observation that pixels with similar appearance in the reference view share similar depth, we use the same context feature F_c to serve as context-guided information for implicitly propagating depth details from high-confidence to low-confidence regions via attention mechanism. To mitigate the computational burden arising from attention weights based on the dot product of query Q and key K, which grows quadratically with input image resolution, we adopt linear attention [17] and propose the Context-Guided Global Aggregation (CGA) module to propagate depth information, as shown in Figure 4. In this module, we apply 2D absolute positional embedding to F_c, enabling attention weights to capture self-similarity and absolute position of the context feature, and then reshape and project them into query $\mathrm{Q} \in \mathbb{R}^{N\times C}$ and key $\mathrm{K} \in \mathbb{R}^{N\times C}$ matrices ($N = H \times W$) using a fully connected layer. Concurrently, we reshape and project the encoded cost volume $\mathrm{C} \in \mathbb{R}^{C\times H\times W}$, which has three channels mentioned in Sec III-C, into a value matrix $\mathrm{V} \in \mathbb{R}^{N\times C}$ ($N = H \times W$) with another fully connected layer. The context-guided global aggregated volume is defined as follows:

$$\mathrm{C}_L = \alpha\Phi(\mathrm{Q})(\Phi(\mathrm{K}^\top)\mathrm{V}) + \mathrm{V}, \tag{7}$$

where α is a learned scalar parameter initialized to zero, $\Phi(\cdot) = elu(\cdot) + 1$, and elu$(\cdot)$ is the exponential linear unit activation function, which avoids zero gradients for negative inputs, unlike relu$(\cdot)$. The dimension of C_L is reshaped to $[C, H, W]$. Linear computational complexity makes it feasible to process high-resolution images on mainstream GPU devices.

C. Network Architecture

We integrate the CWF and CGA modules into Effi-MVS [7], proposing a new network named ACP-MVS, as shown in Figure 2.

Feature Extraction. Similar to [27], we employ a Feature Pyramid Network to extract multi-scale features from the reference image and $N-1$ source images with resolution $3 \times H \times W$ at three scale stages ($k = 0, 1, 2$), producing features of size $\frac{H}{2^{(3-k)}} \times \frac{W}{2^{(3-k)}}$. Similarly, we use the context extractor which is constructed in the same way as the feature extractor to extract multi-scale context features from the reference image for the proposed modules and GRU-based regularization.

CWF and CGA Module. Accurate depth maps at the coarse stage are crucial for reducing cumulative errors in the cascade framework. To address this, we use CWF and CGA to refine the coarse-stage depth map. First, CWF constructs a context-stimulated cost volume by sparsely sampling depth hypotheses over a wide inverse depth range. A lightweight 3D regularization network aggregates this cost volume into a probability volume, and D_{init} is regressed via soft-argmin. Next, a compact cost volume is constructed for CGA to minimize memory usage. Specifically, we first sample a limited number of depth hypotheses within a narrow inverse depth range based on D_{init}. Details on the number of depth hypotheses and depth hypothesis intervals are provided in Sec IV-B. Based on the latest depth hypotheses, we construct a cost volume:

$$\mathrm{C} = \frac{1}{N-1}\sum_{i=1}^{N-1}(\mathrm{V}_{ref} - \mathrm{V}_i)^2, \tag{8}$$

where V_{ref} and V_i denote the reference volume and the source volume of view i, respectively.

Following [7], C is fused with D_{init} by a 2D CNN to produce the encoded volume $\mathrm{C}_e \in \mathbb{R}^{C\times D\times H\times W}$. CGA processes C_e to generate the context-guided global aggregated volume C_L. Details are as follows:

$$\mathrm{C}_e = \mathrm{concat}[\mathrm{Conv}(\mathrm{C}, \mathrm{D}_{init}), \mathrm{D}_{init}] \tag{9}$$

$$\mathrm{C}_L = \mathrm{CGA}(\mathrm{C}_e, \mathrm{F}_{context}) \tag{10}$$

Finally, C_L, C_e and reference-image context features are concatenated along the channel dimension as input to the GRU-based iterative regularization. The concatenation operation enables the network to adaptively choose or merge local matching information and context-guided global details, enhancing the decoding of more accurate depth maps from the coarse-stage cost volume.

GRU-based Regularization. We employ a multi-stage GRU-based iterative architecture to leverage multi-scale information. The GRU-based regularization module updates the depth map T times at each stage k. In each iteration t, this module takes reference-image context features and the current estimated depth, and outputs a delta depth Δd, which is added to the current depth map to obtain an updated depth map used as the input for the next iteration $t+1$. The depth map in the final iteration of each stage is upsampled via a convex upsampler [28]. The upsampled depth map serves as the initial map for the next stage. Subsequently, the network resamples depth hypotheses based on the upsampled depth and constructs a new cost volume until the final depth map matches the input image resolution.

Loss Function. Following previous work [1], [8], [9], we use ℓ_1 loss for depth regression between predictions and

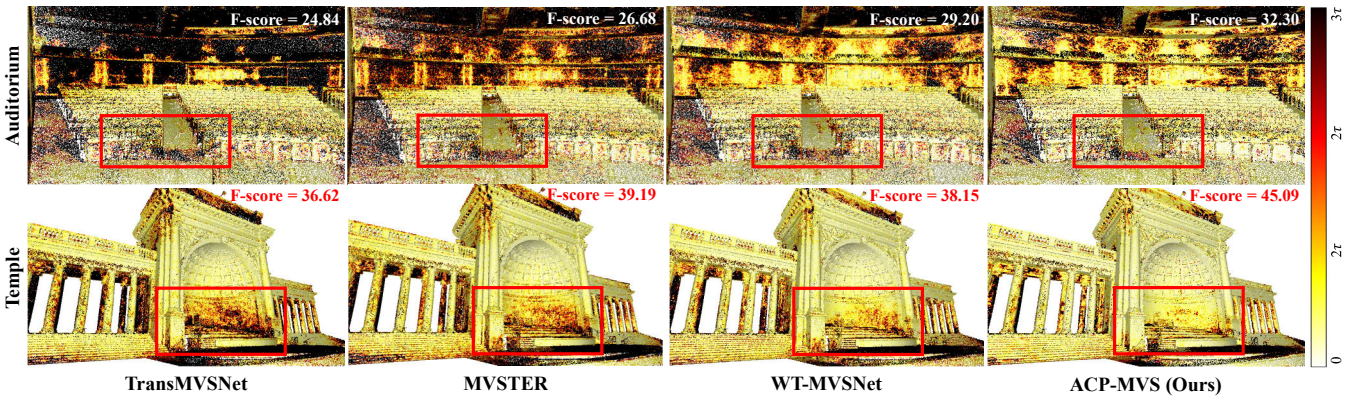


Fig. 5. **Comparison of reconstruction results with state-of-the-art attention-based methods [13], [15], [6] on the Tanks and Temples benchmark.** τ denotes the officially determined scene-relevant distance threshold, and darker regions correspond to higher errors. The first and the second rows show the Recall for the advanced scene of the Auditorium ($\tau = 10mm$) and Temple ($\tau = 15mm$).

ground truth. Our network produces depth maps from the CWF module and multi-stage GRU-based regularization, so the final loss is defined as:

$$Loss = L_{\text{CWF}} + \sum_{k=0}^{2} \sum_{i=1}^{T_k+1} \gamma_i^k L_i^k, \tag{11}$$

where L_{CWF} is the loss from the CWF module. T_k is the number of iterations at stage k. $\{L_i^k \mid i = 1 \dots T_k{+}1\}$ include the loss of T_k iterations and an upsampled depth map at stage k, and γ_i^k is the corresponding weight.

IV. Experiments

A. Datasets and Evaluation Metrics

Datasets. DTU [4] is a substantial indoor MVS dataset with over 100 scenes under seven different lighting conditions. It is split into training, validation, and evaluation sets. BlendedMVS [29] is a large-scale dataset with 17,000 samples from 113 scenes, split into training and validation sets. Tanks and Temples [5] includes a range of realistic outdoor and indoor scenes, divided into intermediate and advanced subsets.

Evaluation Metrics. To evaluate point cloud quality, we use distance-based accuracy and completeness for DTU and the accuracy and completeness of the percentage metric for Tanks and Temples. For overall evaluation, we calculate the average of the accuracy and completeness for DTU and the F1 score for Tanks and Temples.

B. Implementation Details

Following the common practice, we train ACP-MVS on DTU and fine-tune on BlendedMVS. During training, we set the number of input images N = 5 with a resolution of 640×512 for DTU and N = 7 with a resolution of 768×576 for BlendedMVS. For the CWF module, we set the number of depth hypotheses at 48 for DTU and 96 for BlendedMVS. For the CGA and GRU-based regularization module, we keep the number of depth hypotheses at 4 for all stages. The inverse depth interval I_m is defined as:

$$I_m = (\frac{1}{d_{min}} - \frac{1}{d_{max}})/Z \tag{12}$$

We set Z as 384 for DTU and 768 for BlendedMVS, along with depth hypothesis intervals at stages $k = 0, 1, 2$ set to $4I_m$, $2I_m$, and I_m, respectively. For GRU-based regularization, the iteration number T_k is set to 3 for all stages. ACP-MVS is implemented in PyTorch and trained with AdamW under the OneCycleLR for 20 epochs with a learning rate of 0.001 and a batch size of 12 on two NVIDIA GeForce RTX 3090 GPUs. For depth filtering and fusion, we use the improved filtering algorithm proposed in [7] for DTU, and adopt the dynamic checking fusion strategy [34] for Tanks and Temples.

C. Time, Memory and Reconstruction Quality

To demonstrate the high efficiency of our method, we provide a comparative assessment of ACP-MVS alongside open-source learning-based MVS methods [6], [13], [14], [18], [20], [25] on the DTU and Tanks and Temples benchmark, as shown in Table I. For fair comparisons, we use a fixed input size of 1600×1152 to evaluate memory and inference time on a single NVIDIA GeForce RTX 3090 GPU. ACP-MVS's GPU memory usage and inference time are normalized to 100% as the baseline.

Comparison with attention-based methods. Since ACP-MVS employs the attention mechanism, we compare our approach with attention-based methods. Compared to TransMVSNet [13], ACP-MVS reduces GPU memory by 73.47% and runtime by 83.73%, while exhibiting significantly enhanced reconstruction performance on DTU and Tanks and Temples. Compared to efficient MVSTER [6], ACP-MVS improves by 11.16% and 6.20% on the Tanks and Temples advanced and intermediate subsets respectively, while reducing memory by 46.52% and runtime by 14.53%. When compared to MVSFormer [14], ACP-MVS reduces memory by 66.89% and runtime by 59.18% respectively, with comparable performance. Notably, MVSFormer employs a sophisticated pre-trained strategy involving vision transformer fine-tuning on 2 V100 GPUs and requires 20 views for testing to achieve the current performance. In contrast, our network is device-friendly, avoids complex training strategies, and delivers remarkable results cost-effectively.

TABLE I

COMPARISON OF PERFORMANCE, GPU MEMORY USAGE, AND INFERENCE TIME WITH ATTENTION-BASED AND EFFICIENT METHODS ON DTU AND TANKS AND TEMPLES(TAT). '∗' DENOTES METHODS TRAINED SOLELY ON DTU. 'ITERS' REPRESENTS THE NUMBER OF GRU ITERATIONS AT EACH STAGE. MEMORY AND RUNTIME OF ACP-MVS (ITERS: 3 3 3) ARE NORMALIZED TO 100% AS THE BASELINE. **BOLD** REPRESENTS THE BEST AND UNDERLINED REPRESENTS THE SECOND-BEST.

	Method	Memory (%)	Time (%)	DTU (mm)	TAT (Advanced)	TAT (Intermediate)
Att.	TransMVSNet [13]	377%	615%	0.305	37.00	63.52
	MVSTER [6]	187%	117%	0.303	37.53	60.92
	MVSFormer [14]	302%	245%	**0.289**	40.87	**66.37**
Effi.	CasMVSNet* [18]	399%	230%	0.355	31.12	56.84
	Patchmatchnet* [20]	145%	130%	0.352	32.31	53.15
	IterMVS [25]	**73%**	95%	0.363	34.17	56.94
Ours* (Iters: 3 3 3)		100%	100%	0.300	37.41	59.81
Ours (Iters: 1 1 1)		100%	**68%**	0.306	40.38	63.48
Ours (Iters: 3 3 3)		100%	100%	0.300	**41.72**	64.70

TABLE II

QUANTITATIVE RESULTS ON TANKS AND TEMPLES. METHODS ARE CATEGORIZED INTO THREE GROUPS: TRADITIONAL METHODS, METHODS TRAINED ON DTU, AND FINE-TUNED ON BLENDEDMVS. **BOLD** REPRESENTS THE BEST AND UNDERLINED REPRESENTS THE SECOND-BEST.

	Method	Advanced							Intermediate								
		Mean↑	Aud.	Bal.	Cou.	Mus.	Pal.	Tem.	Mean↑	Fam.	Fra.	Hor.	Lig.	M60.	Pan.	Pla.	Tra.
Tra.	COLMAP [2]	27.24	16.02	25.23	34.70	41.51	18.05	27.94	42.14	50.41	22.25	25.63	56.43	44.83	46.97	48.53	42.04
	ACMM [30]	34.02	23.41	32.91	41.17	48.13	23.87	34.60	57.27	69.24	51.45	46.97	63.2	55.07	57.64	60.08	54.48
DTU	CasMVSNet [18]	31.12	19.81	38.46	29.10	43.87	27.36	28.11	56.42	76.36	58.45	46.20	55.53	56.11	54.02	58.17	46.56
	PatchmatchNet [20]	32.31	23.69	37.73	30.04	41.80	28.31	32.29	53.15	66.99	52.64	43.24	54.87	52.87	49.54	54.21	50.81
	Effi-MVS [7]	34.39	20.22	42.39	33.73	45.08	29.81	35.09	56.88	72.21	51.02	51.78	58.63	58.71	56.21	57.07	49.38
	ACP-MVS (Ours)	37.41	23.29	43.84	37.48	48.39	32.44	39.05	59.81	76.24	55.69	53.01	62.49	60.32	56.79	58.25	55.70
BlendedMVS	IterMVS [25]	34.17	25.90	38.41	31.16	44.83	29.59	35.15	56.94	76.12	55.8	50.53	56.05	57.68	52.62	55.70	50.99
	TransMVSNet [13]	37.00	24.84	44.59	34.77	46.49	34.69	36.62	63.52	80.92	65.83	56.94	62.54	63.06	60.00	60.20	58.67
	GBi-Net [24]	37.32	29.77	42.12	36.3	47.69	31.11	36.93	61.42	79.77	67.69	51.81	61.25	60.37	55.87	60.67	53.89
	MVSTER [6]	37.53	26.68	42.14	35.65	49.37	32.16	39.19	60.92	80.21	63.51	52.30	61.38	61.47	58.16	58.98	51.38
	UniMVSNet [31]	38.96	28.33	44.36	39.74	52.89	33.80	34.63	64.36	81.20	66.43	53.11	63.46	66.09	64.84	62.23	57.53
	MVSFormer [14]	40.87	28.22	46.75	39.30	52.88	35.16	42.95	66.37	82.06	69.34	60.49	68.61	65.67	64.08	61.23	59.53
	GeoMVSNet [32]	41.52	30.23	46.53	39.98	53.05	35.98	43.34	65.89	81.64	67.53	55.78	68.02	65.49	67.19	63.27	58.22
	GoMVS [33]	**43.07**	35.52	47.15	42.52	52.08	36.34	44.82	**66.44**	82.68	69.23	69.19	63.56	65.13	62.10	58.81	60.80
	ACP-MVS (Ours)	41.72	32.30	46.53	39.35	51.23	35.81	45.09	64.70	80.89	68.73	55.97	66.16	63.45	61.84	61.91	58.65

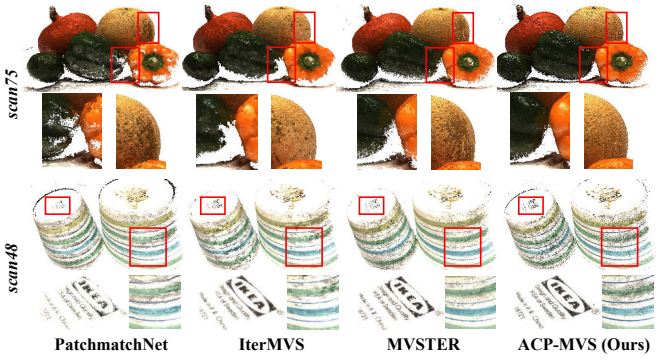


Fig. 6. **Comparison of reconstruction results with state-of-the-art efficient methods [20], [25], [6] on the DTU evaluation set.** Our method performs well in untextured and low-texture regions.

Comparison with efficient methods. We also compare with multistage methods tailored for both memory and inference efficiency. Compared to Patchmatchnet [20] and CasMVSNet [18], ACP-MVS has achieved a significant increase in reconstruction performance while maintaining high memory and runtime efficiency. In particular, our method with fewer GRU iterations outperforms IterMVS [25], while maintaining comparable efficiency, demonstrating strong generalization. ACP-MVS adjusts the number of iterations based on the actual application needs flexibly.

D. Benchmark Performance

Results on Tanks and Temples. To evaluate generalization, we test ACP-MVS on the Tanks and Temples benchmark. We utilize 11 views at 1920×1024 resolution. For a fair comparison, we test two models: one trained solely on DTU and another fine-tuned on BlendedMVS. Quantitative results for intermediate and advanced datasets are shown in Table II. ACP-MVS achieves state-of-the-art performance. Specifically, our DTU-trained model outperforms all learning-based methods trained only on DTU. We rank second in the advanced subset among all published works. Compared to the intermediate subset, the advanced subset presents more challenges, such as weaker illumination, numerous surfaces with nearly uniform appearances, and other complicating factors. This demonstrates the robustness and generalization capabilities of ACP-MVS under extensive and challenging scenarios. Furthermore, our method achieves higher F1 scores compared to state-of-the-art efficient methods [25], [7], [20], [6], further confirming its superiority. Figure 5 illustrates point cloud error comparisons, highlighting ACP-MVS's enhanced recall.

Results on DTU. We evaluate ACP-MVS on the evaluation set of the DTU dataset using the model only trained on the DTU training set. We use 5 views at 1600×1152 resolution. Qualitative results are shown in Figure 6. We compare

TABLE III

QUANTITATIVE RESULTS ON DTU. THE METHODS ARE CATEGORIZED INTO THREE GROUPS: TRADITIONAL METHODS, CONVOLUTION-BASED METHODS, AND ATTENTION-BASED METHODS.

	Method	Acc. (mm)	Comp. (mm)	Overall (mm)	Mem. (MB)	Time (s)
Tra.	COLMAP [2]	0.411	0.657	0.534	-	-
	Gipuma [1]	0.283	0.873	0.578	-	-
Con.	Vis-MVSNet [35]	0.369	0.361	0.365	4775	1.121
	IterMVS [25]	0.373	0.354	0.363	845	0.189
	CasMVSNet [18]	0.325	0.385	0.355	4586	0.456
	PatchmatchNet [20]	0.427	0.277	0.352	1670	0.258
	IGEV-MVS [36]	0.331	0.316	0.324	6895	3.130
	Effi-MVS [7]	0.321	0.313	0.317	1001	0.185
	UniMVSNet [31]	0.352	0.278	0.315	6120	0.648
	GBiNet [24]	0.312	0.293	0.303	2130	0.671
	GeoMVSNet [32]	0.331	0.259	0.295	4734	0.344
	GoMVS [33]	0.347	0.227	0.287	-	-
Att.	CostFormer [16]	0.301	0.322	0.312	-	-
	TransMVSNet [13]	0.321	0.289	0.305	4337	1.218
	MVSTER [6]	0.340	0.266	0.303	2152	0.232
	WT-MVSNet [15]	0.309	0.281	0.295	-	-
	MVSFormer [14]	0.327	0.251	0.289	3471	0.486
	ACP-MVS (Ours)	0.315	0.285	0.300	1149	0.198

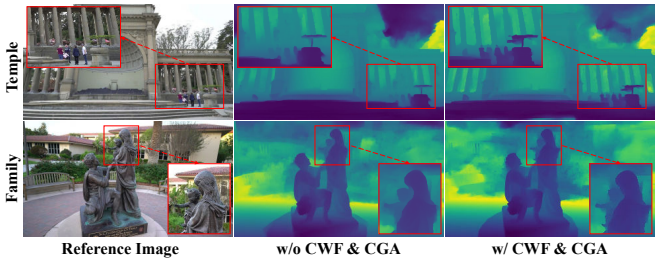


Fig. 7. **Qualitative comparisons of estimated depth maps on the Tanks and Temples benchmark.** CWF and CGA significantly improve performance in both distant and close-range scenarios.

ACP-MVS with state-of-the-art efficient methods, focusing on the point clouds for $scan48$ and $scan75$, which have reflections and low-texture regions. As depicted, our method excels in accurately recovering point clouds in challenging regions. Detailed quantitative results on DTU are presented in Table III. Our method achieves state-of-the-art results and is the most efficient among methods whose Overall metric is less than 0.315 mm, in terms of runtime and memory usage.

E. Ablation Study

We conduct ablation studies to assess the impact of CWF and CGA, as shown in Table IV and Figure 8. All models in this experiment are trained and tested on DTU. We set Effi-MVS [7] as the baseline method. Qualitative results in Figure 7 further validate the effectiveness of CWF and CGA.

Context-stimulated weight fusion. In Table IV, 'CON' denotes per-view weights generated by employing a 3D convolution followed by a batch normalization to two-view cost volumes. Both weight generation methods show advantages on DTU, but CWF notably delivers superior results with minimal time and memory overhead. The context-stimulated weights effectively capture geometric information of the reference image. This helps the network mitigate the impact of invalid pixels while enhancing pixels with crucial context.

TABLE IV

ABLATION RESULTS ON THE DTU EVALUATION SET. THE SETTINGS EMPLOYED IN ACP-MVS ARE INDICATED BY UNDERLINED.

Experiment	Variations	Overall (mm)	Mem. (MB)	Time (s)
baseline	-	0.317	1001	0.185
Fusion	CON	0.314	1055	0.191
	CWF	0.309	1060	0.192
Aggregation	MA	0.305	7075	0.227
	CGA	0.307	1040	0.189
ACP-MVS (Ours)	CWF + CGA	0.300	1149	0.198

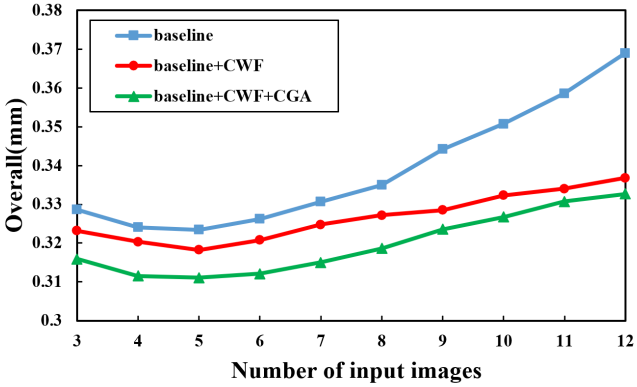


Fig. 8. Comparison of reconstruction results from different models with varying input numbers.

Context-guided global aggregation. In Table IV, 'MA' denotes that the baseline method only employs multi-head self-attention to compute context-guided information for global aggregation without additional setups. Both methods improve the reconstruction performance, highlighting the significance of our propagation approach. Multi-head attention improves results slightly, but significantly increases memory overhead. Consequently, we select CGA with minimal memory overhead as the final configuration.

varying input numbers. Current datasets lack masks for challenging regions, hindering quantitative evaluation of these areas. To address this, we increase the number of input images and validate our proposed modules by tackling the challenges from perspective changes. For a fair comparison, we train models using only the two best adjacent views. As depicted in Figure 8, the performance of our baseline deteriorates significantly with more input views. This is because the original training strategy focuses on the two best adjacent views, which typically have high pixel-wise visibility probability. Consequently, the trained network tends to overfit matching regions, leading to poor discrimination of ill-posed areas. Ignoring pixel-wise mismatches introduces noise as the number of views increases, severely degrading performance. However, CWF enhances reliable pixels with larger weights and suppresses irrelevant or noisy information, mitigating issues from increased views. Even with 12 input views, the network's performance remains stable. Furthermore, CGA uses attention-based context to propagate enhanced information, further improving performance.

V. CONCLUSIONS

In this paper, we introduce an efficient context-perception network known as ACP-MVS, which adaptively incorporates

context information via lightweight attention mechanism. Specifically, ACP-MVS utilizes the CWF module to generate context-stimulated weights for cost volume fusion. Additionally, we introduce the CGA module to propagate enhanced matching information. These two collaborating modules enhance the performance of ACP-MVS without significantly increasing computational and time costs. Our approach achieves state-of-the-art performance efficiently on both DTU and Tanks and Temples benchmark, providing practical benefits for high-resolution MVS applications.

REFERENCES

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

[2] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4104–4113.

[3] Z. Yuan, F. Lang, J. Deng, H. Luo, and X. Yang, "Voxel-svio: Stereo visual-inertial odometry based on voxel map," *IEEE Robotics and Automation Letters*, vol. 10, no. 6, pp. 6352–6359, 2025.

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

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

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

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

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

[9] J. Yang, W. Mao, J. M. Alvarez, and M. Liu, "Cost volume pyramid based depth inference for multi-view stereo," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4877–4886.

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

[11] Z. Wei, Q. Zhu, C. Min, Y. Chen, and G. Wang, "Aa-rmvsnet: Adaptive aggregation recurrent multi-view stereo network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6187–6196.

[12] H. Yi, Z. Wei, M. Ding, R. Zhang, Y. Chen, G. Wang, and Y.-W. Tai, "Pyramid multi-view stereo net with self-adaptive view aggregation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*. Springer, 2020, pp. 766–782.

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

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

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

[16] W. Chen, H. Xu, Z. Zhou, Y. Liu, B. Sun, W. Kang, and X. Xie, "Costformer: Cost transformer for cost aggregation in multi-view stereo," *arXiv preprint arXiv:2305.10320*, 2023.

[17] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, "Transformers are rnns: Fast autoregressive transformers with linear attention," in *International conference on machine learning*. PMLR, 2020, pp. 5156–5165.

[18] X. Gu, Z. Fan, S. Zhu, Z. Dai, F. Tan, and P. Tan, "Cascade cost volume for high-resolution multi-view stereo and stereo matching," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2495–2504.

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

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

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

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

[24] Z. Mi, C. Di, and D. Xu, "Generalized binary search network for highly-efficient multi-view stereo," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 991–13 000.

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

[26] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," *Advances in neural information processing systems*, vol. 28, 2015.

[27] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[28] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *European conference on computer vision*. Springer, 2020, pp. 402–419.

[29] Y. Yao, Z. Luo, S. Li, J. Zhang, Y. Ren, L. Zhou, T. Fang, and L. Quan, "Blendedmvs: A large-scale dataset for generalized multi-view stereo networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 1790–1799.

[30] Q. Xu and W. Tao, "Multi-scale geometric consistency guided multi-view stereo," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5483–5492.

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

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

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

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

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

[36] G. Xu, X. Wang, X. Ding, and X. Yang, "Iterative geometry encoding volume for stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 919–21 928.