TransMVSNet: Global Context-aware Multi-view Stereo Network with Transformers

Yikang Ding[1,2*] Wentao Yuan[1,3*] Qingtian Zhu[1] Haotian Zhang[1]
Xiangyue Liu[1] Yuanjiang Wang[1†] Xiao Liu[1‡]

[1] Megvii Research [2] Tsinghua University [3] Peking University

Abstract

In this paper, we present TransMVSNet, based on our exploration of feature matching in multi-view stereo (MVS). We analogize MVS back to its nature of a feature matching task and therefore propose a powerful Feature Matching Transformer (FMT) to leverage intra- (self-) and inter- (cross-) attention to aggregate long-range context information within and across images. To facilitate a better adaptation of the FMT, we leverage an Adaptive Receptive Field (ARF) module to ensure a smooth transit in scopes of features and bridge different stages with a feature pathway to pass transformed features and gradients across different scales. In addition, we apply pair-wise feature correlation to measure similarity between features, and adopt ambiguity-reducing focal loss to strengthen the supervision. To the best of our knowledge, TransMVSNet is the first attempt to leverage Transformer into the task of MVS. As a result, our method achieves state-of-the-art performance on DTU dataset, Tanks and Temples benchmark, and BlendedMVS dataset. Code is available at https://github.com/MegviiRobot/TransMVSNet.

1. Introduction

Multi-view stereo (MVS) aims to recover the dense 3D presentation with a series of calibrated images, which is an important task of computer vision. Learning-based MVS networks [10, 31, 32] have achieved remarkable progress in terms of reconstruction quality and efficiency. Typically, a MVS network extracts image features by a CNN and constructs cost volume via plane sweep algorithm [5] in which source images are warped to the reference view. This cost volume is regularized afterwards to estimate the final depth.

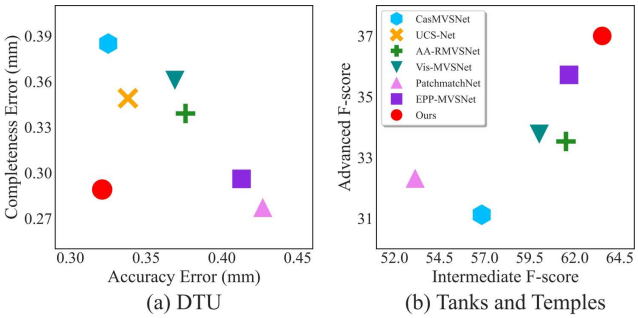

Figure 1. Comparison with state-of-the-art learning-based MVS methods [3, 10, 19, 26, 27, 34] on DTU dataset [1] (**lower is better**) and Tanks and Temples benchmark [12] (**higher is better**).

The nature of MVS is a one-to-many feature matching task, in which each pixel of the reference image is supposed to search along the epipolar line in all warped source images and find an optimal depth with the lowest matching cost. Some recent studies [22, 24] have proven the importance of long-range global context in feature matching tasks. However, given the aforementioned MVS pipeline, there are two main problems. (a) Local features are well captured by convolutions. The locality of convolved features prevents the perception of global context information, which is essential for robust depth estimation at challenging regions in MVS, *e.g.* poor texture, repetitive patterns, and non-Lambertian surfaces. (b) Besides, when computing matching costs, the features to be compared are simply extracted respectively from each image itself, which is to say, potential inter-image correspondences are not taken into consideration.

Recently, Transformer [25], which is initially proposed for natural language processing, has drawn considerable attention from the computer vision community for their great performance on vision tasks. Since Transformer utilizes the mechanism of attention and positional encoding for context aggregation, rather than convolutions, it is capable of perceiving global and positionally relevant context information

*Equal Contribution. This work is done by the author as interns at Megvii Research.
†Project lead.
‡Corresponding author (liuxiao@foxmail.com).

in the true sense.

To this end, we propose a novel end-to-end deep neural network, namely TransMVSNet, to which a powerful Feature Matching Transformer (FMT) is leveraged to strengthen long-range global context aggregation within and between images. To better adapt FMT into an end-to-end learning-based MVS pipeline, we introduce an Adaptive Receptive Field (ARF) module to ensure a smooth transition from locally aggregated features by CNN to features with a global receptive field by FMT. In order to lower runtime memory requirements and train FMT with supervision from high-resolution depth maps, we bridge different scales with a transformed feature pathway. We apply pair-wise feature correlation to measure the similarity between the reference feature map and each of its source feature maps. Afterwards, we follow the coarse-to-fine volume regularization pattern [10] and adopt focal loss [16], which better handles samples with ambiguous prediction, to end-to-end train the network.

Thanks to the global context-aware information within and between views, TransMVSNet achieves significant improvement in reconstruction accuracy and completeness simultaneously on DTU dataset [1] (as shown in Fig. 1(a)). Moreover, the overwhelming performance of TransMVSNet can be generalized to more complex scenes, *e.g*. the intermediate and advanced set of Tanks and Temples benchmark [12] (as shown in Fig. 1(b)). To the best of our knowledge, it is the first attempt that takes advantage of Transformer in the task of MVS. Consequently, extensive experiments indicate that our method achieves state-of-the-art performance. We also conduct ablation experiments to demonstrate the effectiveness of each proposed module. Our main contributions are three-fold as follows.

- We propose a novel end-to-end deep neural network based on a Feature Matching Transformer (FMT), namely TransMVSNet, for robust long-range global context aggregation within and across images.
- To better adapt FMT into an end-to-end MVS pipeline, we introduce an ARF module to adaptively adjust the receptive fields of convolved features and apply ambiguity-aware focal loss for training.
- Our method achieves state-of-the-art results on DTU dataset, Tanks and Temples benchmark, and BlendedMVS dataset.

2. Related Work

2.1. Learning-based MVS

In the modern deep era, learning-based methods have been introduced to the task of MVS for better reconstruction accuracy and completeness. MVSNet [31] encodes camera parameters via differentiable homography to build 3D cost volumes, and decouples the MVS task to a per-view depth map estimation task. However, the memory and computation costs are quite expensive due to its 3D U-Net architecture for cost volume regularization. To alleviate this problem, several networks have been proposed and can be categorized into RNN-based recurrent methods [27, 29, 32] and coarse-to-fine multi-stage methods [3, 10, 30, 34], according to regularization patterns, Recurrent methods regularize the 3D cost volumes recurrently, and adopt RNNs to pass features between different depth hypotheses. Since recurrent methods trade time for space, they are capable of handling images with large resolution but slow in terms of inference speed. Multi-stage methods predict a coarse depth map initially and narrow down the target depth range at a larger resolution based on the previous prediction. Coarse-to-fine methods are able to infer quickly while keeping a relatively small memory consumption.

Though learning-based MVS methods have achieved promising results, there are still challenging problems remaining, *e.g*. robust estimation at non-Lambertian and low-texture regions or severely occluded areas.

2.2. Transformer for Feature Matching

Transformer [25] has been widely used in natural language processing due to its effectiveness and efficiency, and has drawn increasing attention from the computer vision community recently [2, 8, 17, 20, 21]. Considering Transformer's natural superiority to capture global context information by leveraging attention, its ideology has been utilized in the task of feature matching.

SuperGlue [22] utilizes self- and cross-attention in the task of sparse feature matching, leveraging both spatial relationships and visual appearance of the keypoints. SuperGlue achieves impressive performance and becomes the new state of the art. LoFTR [24] establishes accurate dense matches with Transformers in a coarse-to-fine manner. By interleaving the self- and cross-attention layers multiple times, LoFTR learns densely arranged and globally consented matching priors in ground-truth matches. STTR [14] models the task of stereo depth estimation from a sequence-to-sequence matching perspective. Transformers with alternating self- and cross-attention along intra- and inter-epipolar line are adopted to capture long-range associations between feature descriptors.

3. Methodology

Given a reference image $\mathbf{I}_0 \in \mathbb{R}^{H \times W \times 3}$ and its neighboring images $\{\mathbf{I}_i\}_{i=1}^{N-1}$, as well as their respective camera intrinsics and extrinsics, our method predicts a depth map aligned with $\mathbf{I}_0$. Depth maps of all images are then filtered and fused to obtain the reconstructed dense point cloud.

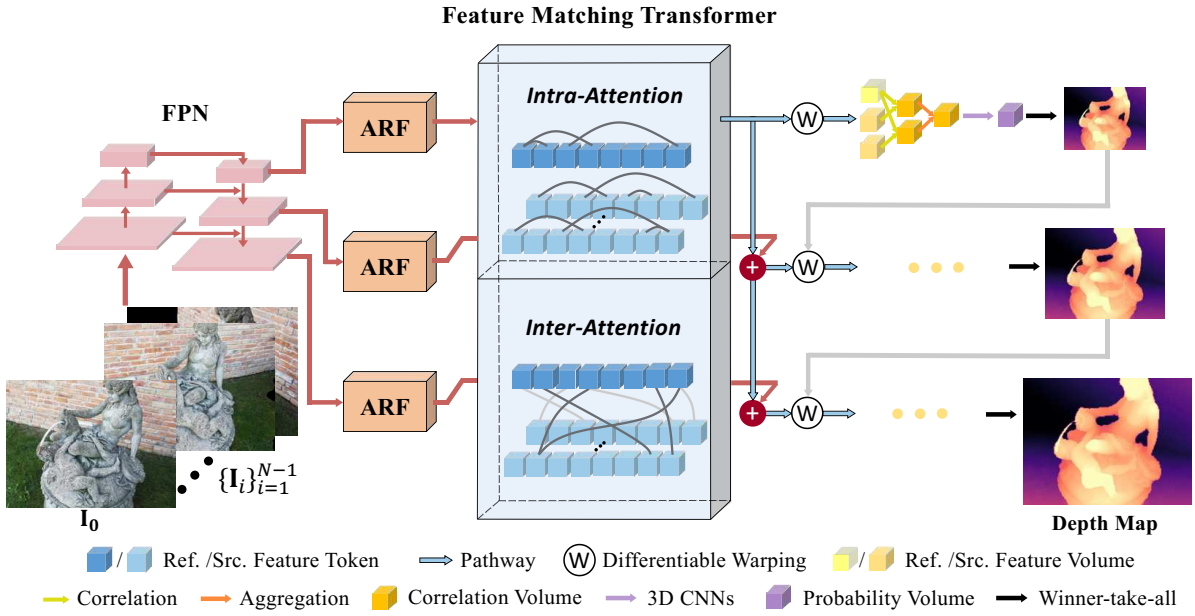

Figure 2. TransMVSNet architecture. TransMVSNet extracts basic features by FPN and introduces ARF modules (Sec. 3.4) to ensure the transit from FPN to Transformer. In the FMT (Sec. 3.2), intra-attention is performed to aggregate global context within images and inter-attention helps feature searching and matching across images. A transformed feature pathway (Sec. 3.3) is connected to pass low-resolution features to higher resolutions and enable back-propagating gradients of all scales to go through FMT. We then apply pixel-wise feature correlation to generate the correlation volumes (Sec. 3.5), which are regularized with a coarse-to-fine pattern.

3.1. Network Overview

The overall architecture of our TransMVSNet is illustrated in Fig. 2. TransMVSNet first applies a Feature Pyramid Network (FPN) [15] to extract multi-scale deep image features at three coarse-to-fine levels of resolution. Before handing these features to Transformer, we use the Adaptive Receptive Field (ARF) module, described in Sec. 3.4, to refine the local feature extraction and ensure a smooth transit to Transformer. To leverage global context information within and between reference and source images, we adopt the Feature Matching Transformer (FMT) to perform intra- and inter-attention. The technical details of FMT are introduced in Sec. 3.2. To effectively and efficiently propagate transformed features from a low resolution to a higher and make FMT trained with gradients from all scales, we connect all resolutions with a feature pathway described in Sec. 3.3. To be described in Sec. 3.5, for feature maps of $N \times H' \times W' \times F$ processed by FMT, we build a correlation volume of $H' \times W' \times D' \times 1$ for the following regularization by 3D CNNs. H', W' and F denote the height, width and channels of feature maps at current stage, N denotes the number of views and D' denotes the corresponding number of depth hypotheses. After obtaining the regularized probability volume, we take the strategy of winner-take-all to determine the final prediction. We apply focal loss with enhanced punishment at ambiguous areas, as described in Sec. 3.6, to train TransMVSNet end-to-end.

3.2. Feature Matching Transformer (FMT)

For most cases, learning-based MVS networks construct cost volumes directly from extracted features, ignoring global context information and inter-image feature interaction, which have been proven to be important for improving prediction quality and reducing uncertainty of matching, especially for low-textured regions and repetitive patterns. Aforementioned Transformer-based matching methods handle the problem of feature matching between two views. For MVS, whose nature is a one-to-many matching task, we present a Feature Matching Transformer (FMT), specially customized for MVS. Sec. 3.2.1 introduces the preliminaries of attention; Sec. 3.2.2 further describes the attention mechanism used in the proposed FMT, especially its customization dedicated to MVS; Sec. 3.2.3 demonstrates the design of FMT module as a whole.

3.2.1 Preliminaries

Scaled dot-product attention Analogous to the conventions in information retrieval, features are grouped as query $\mathbf{Q}$, key $\mathbf{K}$ and value $\mathbf{V}$. $\mathbf{Q}$ retrieves relevant information from $\mathbf{V}$ according to the attention weight obtained from the dot product of $\mathbf{Q}$ and $\mathbf{K}$ corresponding to each $\mathbf{V}$. The attention layer is formally denoted as

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax(\mathbf{Q}\mathbf{K}^{\top})\mathbf{V}. \quad (1)$$

The mechanism of attention measures the feature-wise similarity between $\mathbf{Q}$ and $\mathbf{K}$, and retrieves information from $\mathbf{V}$ according to this computed weight. Following the practice in [25], we adopt multi-head attention, which splits the channel of features into N_h groups (number of heads).

Linear attention Multi-head attention [25] calculates the attention from the dot product of $\mathbf{Q}$ and $\mathbf{K}$, leading the computation cost growing quadratically with regard to the length of the input sequence. To lower the computation cost, we follow [11] and use Linear Transformer to compute attention. Linear Transformer replaces the original kernel function with

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \Phi(\mathbf{Q}) \left(\Phi(\mathbf{K}^\top)\mathbf{V}\right), \qquad (2)$$

where $\Phi(\cdot) = elu(\cdot) + 1$ and $elu(\cdot)$ represents the activation function of exponential linear units [4]. Given that the number of channel is far smaller than the length of input sequence, the computation complexity is reduced to linear, making it possible to compute attention upon high-resolution images.

3.2.2 Intra-attention and Inter-attention

When $\mathbf{Q}$ and $\mathbf{K}$ vectors are features from the same image, attention layers retrieve relevant information within the given view. This can essentially be seen as intra-image long-range global context aggregation. In the other case where $\mathbf{Q}$ and $\mathbf{K}$ vectors are from different views, attention layers then capture cross-relationships across these two views and inter-image feature interaction between images is done in this way. In FMT, we perform intra-attention upon both the reference image $\mathbf{I}_0$ and source images $\{\mathbf{I}_i\}_{i=1}^{N-1}$. When computing inter-attention between $\mathbf{I}_0$ and each $\mathbf{I}_i$, only the feature of $\mathbf{I}_i$ is updated.

Here we explain the reason why reference feature $\mathcal{F}_0$ is not supposed to get updated according to source features. When matching the reference image to its neighboring source images, the reference feature should remain invariant to provide an identical target for all source features. The underlying intuition is that the measurement of similarity is only valid given the same image pair, which indicates that the matching confidence is not comparable universally across different pairs. We also conduct ablation experiments on this minor topic and get results to support this assumption. Please refer to the Supplementary Material for more information.

3.2.3 FMT Architecture

Different from a typical one-to-one matching task between two views, MVS tackles a one-to-many matching problem, where context information of all views should be considered simultaneously. To this end, we propose the FMT to capture long-range context information within and across images.

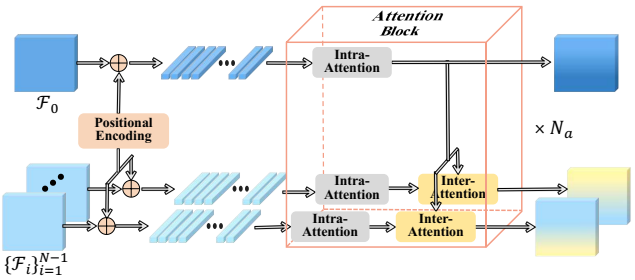

Figure 3. Architecture of the Feature Matching Transformer. FMT performs positional encoding to all features maps and flattens them at the spatial dimension. Then the attention blocks get involved and perform intra- and inter-attention upon features. Note that the number of attention blocks N_a is set as 4 in our implementation.

The architecture of FMT is illustrated in Fig. 3. We follow [22] and add positional encoding, which implicitly enhances positional consistency and makes FMT robust to feature maps with different resolutions. Each view's corresponding flattened feature map $\mathcal{F} \in \mathbb{R}^{H'W' \times F}$ is processed by N_a attention blocks sequentially. Within each attention block (see Fig. 3), the reference feature $\mathcal{F}_0$ and each source feature $\mathcal{F}_i$ firstly compute intra-attention with shared weights, where all features are updated with their respective embedded global context information. Afterwards, the unidirectional inter-attention is performed, with which $\mathcal{F}_i$ is updated according to retrieved information from $\mathcal{F}_0$.

3.3. Transformed Feature Pathway

The Transformer we leverage only performs on feature maps at a rather low resolution since both learning-based MVS and Transformer acquire a massive amount of memory and computation. It remains a problem that how to effectively pass the transformed features from a low resolution to a higher. Besides, we expect the FMT to be trained with supervision from all image scales. We therefore design a transformed feature pathway to fulfill this job. As shown in Fig. 2, feature maps processed by FMT are interpolated to a higher resolution and added to the corresponding raw feature maps at the next image scale.

3.4. Adaptive Receptive Field (ARF) Module

Transformer implicitly encodes global context information into feature maps via positional encoding, which we can roughly perceive as convolution layers with a global receptive field. On the contrary, FPN [15], which is adopted as the basic feature extractor of the proposed network, mainly focuses on the context within a relatively local neighborhood. There is apparently a gap between these two modules in terms of context ranges, which is detrimental to both feature forwarding and end-to-end training.

To this end, we insert an Adaptive Receptive Field Module between FPN and FMT, to adaptively adjust the scope of extracted features. The ARF module is implemented by deformable convolution [6, 35], which learns extra offsets for sampling position and is able to adaptively enlarge the receptive fields according to the local context.

3.5. Correlation Volume Construction

We apply differentiable warping to align all images to the reference view. The warping between a pixel $\mathbf{p}$ at the reference view and its corresponding pixel $\hat{\mathbf{p}}$ at the source view under depth hypothesis d is defined as

$$\hat{\mathbf{p}} = \mathbf{K}[\mathbf{R}(\mathbf{K}_0^{-1}\mathbf{p}d) + \mathbf{t}], \tag{3}$$

where $\mathbf{R}$ and $\mathbf{t}$ denote the rotation and translation between the two views. $\mathbf{K}_0$ and $\mathbf{K}$ are the intrinsic matrices of the reference and source camera. The warped feature maps are bilinearly interpolated to remain the original resolution. By discretizing the known depth space into D depth values, we are able to classify each pixel as one of these values.

Pair-wise feature correlation at position $\mathbf{p}$ is

$$c_i^{(d)}(\mathbf{p}) = < \mathcal{F}_0(\mathbf{p}), \hat{\mathcal{F}}_i^{(d)}(\mathbf{p}) >, \tag{4}$$

where $\hat{\mathcal{F}}_i^{(d)}$ denotes the warped i-th source feature map at depth d. In this way, the channel number is reduced to 1, alleviating subsequent memory consumption at regularization. To aggregate all $N-1$ pair-wise correlation volumes, we consider that each pixel in the height and width dimension of 3D correlation volume has different saliency but is consistent in the depth dimension. We therefore assign a pixel-wise weight map with its maximum correlation in the depth dimension. The aggregated correlation volume is then defined as

$$C^{(d)}(\mathbf{p}) = \sum_{i=1}^{N-1} \max_d\{c_i^{(d)}(\mathbf{p})\} \cdot c_i^{(d)}(\mathbf{p}). \tag{5}$$

3.6. Loss Function

Previous coarse-to-fine attempts [10, 30, 34] mainly adopt $\ell 1$-based depth regression loss that supervises the absolute distance between prediction and ground truth. We instead apply focal loss [16] that treats depth estimation as a classification task to strengthen the one-hot supervision at ambiguous areas. The focal loss at each depth estimation stage is

$$\mathcal{L} = \sum_{\mathbf{p}\in\{\mathbf{p}_v\}} -(1-P^{(\tilde{d})}(\mathbf{p}))^{\gamma} \log\left(P^{(\tilde{d})}(\mathbf{p})\right), \tag{6}$$

where $P^{(d)}(\mathbf{p})$ denotes predicted probability of depth hypothesis d at pixel $\mathbf{p}$ and $\tilde{d}$ represents the depth value closest to the ground truth among all hypotheses. $\{\mathbf{p}_v\}$ represents a subset of pixels with valid ground truth. Specially, focal loss degrades to cross entropy loss when the focusing parameter γ equals 0. Empirically, $\gamma = 2$ fits more complicated scenarios and $\gamma = 0$ can produce good enough results for relatively simple scenarios. Fig. 4 shows the effect of focal loss on boundary regions, where focal loss helps to estimate more accurate boundary than cross entropy loss.

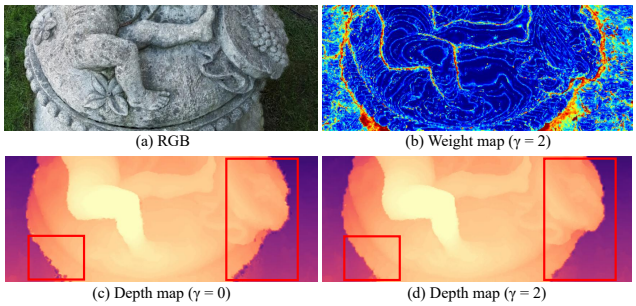

Figure 4. Results visualization of focal loss. (a) Raw image. (b) Focal weight map $(1-P)^{\gamma}$ when the $\gamma = 2$. (c) Depth map when the network is trained with $\gamma = 0$. (d) Depth map when the network is trained with $\gamma = 2$. Focal loss focuses on pixels with low prediction probability, which normally appear in boundary regions.

4. Experiments

4.1. Datasets

DTU [1] is captured under well-controlled laboratory conditions with a fixed camera trajectory and contains 128 scans with 49 views under 7 different lighting conditions. Following the setting of MVSNet [31], we split the dataset into 79 training scans, 18 validation scans, and 22 evaluation scans. BlendedMVS dataset [33] is a large-scale synthetic dataset for multi-view stereo training and contains a variety of objects and scenes. The dataset is split into 106 training scans and 7 validation scans. Tanks and Temples [12] is a public benchmark acquired in realistic conditions. It contains an intermediate subset of 8 scenes and an advanced subset of 6. Different scenes have different scales, surface reflection, and exposure conditions.

4.2. Implementation Details

We implement TransMVSNet with PyTorch and train it on DTU training set [1]. At training phase, we set the number of input images $N = 5$ and image resolution as 512×640. For coarse-to-fine regularization, depth hypotheses are sampled from $425mm$ to $935mm$; the number of plane sweeping depth hypotheses of each stage is respectively 48, 32, and 8; the corresponding depth interval decays by 0.25 and 0.5 from the coarsest stage to the finest stage. The model is trained with Adam for 10 epochs with

Figure 5. Comparison of reconstructed results with state-of-the-art coarse-to-fine methods [3, 10] on DTU evaluation set [1].

an initial learning rate of 0.001, which decays by a factor of 0.5 respectively after 6 and 8 epochs. We set $\gamma = 0$ for training on DTU. The batch size is 1 on 8 NVIDIA RTX 2080Ti GPUs and in total, the training phase takes about 16 hours and occupies 10*GB* memory of each GPU.

For depth filtering and fusion, we follow the dynamic checking strategy proposed in [29], in which both confidence thresholding and geometric consistency are applied.

4.3. Experimental Performance

Evaluation on DTU dataset We evaluate the proposed method on the evaluation set of DTU dataset [1] with official evaluation metrics. We set $N = 5$ and the input resolution as 864×1152 at evaluation phase. As is visualized in Fig. 5, benefiting from the mechanism of intra- and inter-attention in FMT, TransMVSNet is able to yield denser and complete point clouds with more details preserved. Quantitative comparisons are shown in Tab. 1. Accuracy and Completeness are the two official metrics. Accuracy measures the mean absolute point-cloud-to-point-cloud distance from the MVS reconstruction to ground truth, while Completeness measures the opposite. The Overall is the average of Accuracy and Completeness, which indicates the overall performance of models. TransMVSNet achieves competitive performance in Accuracy and Completeness and outperforms all known methods in Overall by a large margin.

Benchmarking on Tanks and Temples To demonstrate the generalization ability of our method, we test our method on Tanks and Temples benchmark [12]. To boost the performance on real-world scenes, we fine-tune TransMVSNet on the training set of the BlendedMVS dataset [33] using the original image resolution (576×768), $N = 5$ and $\gamma = 2$.

For evaluation on Tanks and Temples, the camera parameters, depth ranges, and neighboring view selection are aligned with R-MVSNet [32]. We use images of the original resolution for inference. Quantitative comparisons on Tanks and Temples are shown in Tab. 2 and the metrics are mean F-score. TransMVSNet outperforms all existing learning-based MVS methods on both leaderboards, demonstrating the effectiveness and generalizability of our method. Fig. 6 shows qualitative results on the scene Courtroom of advanced set and Horse of intermediate set. TransMVSNet yields more reliable points at low-textured areas and sophisticated surfaces. Specially, we visualize the process of feature evolution of a pair of views in Fig. 7. In such a typically challenging scene with poor texture and repetitive patterns, FMT manages to capture position-dependent features and aggregate global context within and across different views.

Method	Acc.(mm)	Comp.(mm)	Overall(mm)
Gipuma [9]	**0.283**	0.873	0.578
COLMAP [23]	0.400	0.664	0.532
R-MVSNet [32]	0.385	0.459	0.422
D^2HC-RMVSNet [29]	0.395	0.378	0.386
AA-RMVSNet [27]	0.376	0.339	0.357
Vis-MVSNet [34]	0.369	0.361	0.365
CasMVSNet [10]	0.325	0.385	0.355
UCS-Net [3]	0.338	0.349	<u>0.344</u>
PatchmatchNet [26]	0.427	**0.277**	0.352
EPP-MVSNet [19]	0.413	0.296	0.355
TransMVSNet	<u>0.321</u>	<u>0.289</u>	**0.305**

Table 1. Quantitative results on DTU evaluation set [1] (**lower is better**). **Bold** figures indicate the best and <u>underlined</u> figures indicate the second best. Compared to non-learning methods, RNN-based methods and coarse-to-fine methods, TransMVSNet outperforms all known methods by a large margin.

Evaluation on BlendedMVS dataset Both DTU [1] and Tanks and Temples [12] apply evaluation metrics towards point clouds. We further demonstrate the quality of depth maps, which are the direct outputs by TransMVSNet, on BlendedMVS validation dataset [33]. We set $N = 5$ and image resolution as 512×640, and apply the evaluation

Method	Int.Mean	Family	Francis	Horse	L.H.	M60	Panther	P.G.	Train	Adv.Mean	Auditorium	Ballroom	Courtroom	Museum	Palace	Temple
COLMAP [23]	42.14	50.41	22.25	26.63	56.43	44.83	46.97	48.53	42.04	27.24	16.02	25.23	34.70	41.51	18.05	27.94
ACMM [28]	57.27	69.24	51.45	46.97	63.20	55.07	57.64	60.08	54.48	34.02	23.41	32.91	41.17	48.13	23.87	34.60
DeepC-MVS [13]	59.79	71.91	54.08	42.29	**66.54**	55.77	**67.47**	60.47	**59.83**	34.54	**26.30**	34.66	**43.50**	45.66	23.09	34.00
AttMVS [18]	60.05	73.90	62.58	44.08	64.88	56.08	59.39	**63.42**	56.06	31.93	15.96	27.71	37.99	**52.01**	29.07	28.84
CasMVSNet [10]	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
Vis-MVSNet [34]	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
PatchmatchNet [26]	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
EPP-MVSNet [19]	61.68	77.86	60.54	52.96	62.33	61.69	60.34	62.44	55.30	35.72	21.28	39.74	35.34	49.21	30.00	**38.75**
R-MVSNet [32]	50.55	73.01	54.46	43.42	43.88	46.80	46.69	50.87	45.25	29.55	19.49	31.45	29.99	42.31	22.94	31.10
AA-RMVSNet [27]	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
TransMVSNet	**63.52**	**80.92**	**65.83**	**56.94**	62.54	63.06	60.00	60.20	58.67	**37.00**	24.84	**44.59**	34.77	46.49	**34.69**	36.62

Table 2. Benchmarking results on the Tanks and Temples [12]. The evaluation metric is mean F-score (**higher is better**). **Bold** figures indicate the best and underlined figures indicate the second best. TransMVSNet achieves state-of-the-art performance on both the intermediate and the advanced leaderboards of Tanks and Temples benchmark (Nov. 12, 2021).

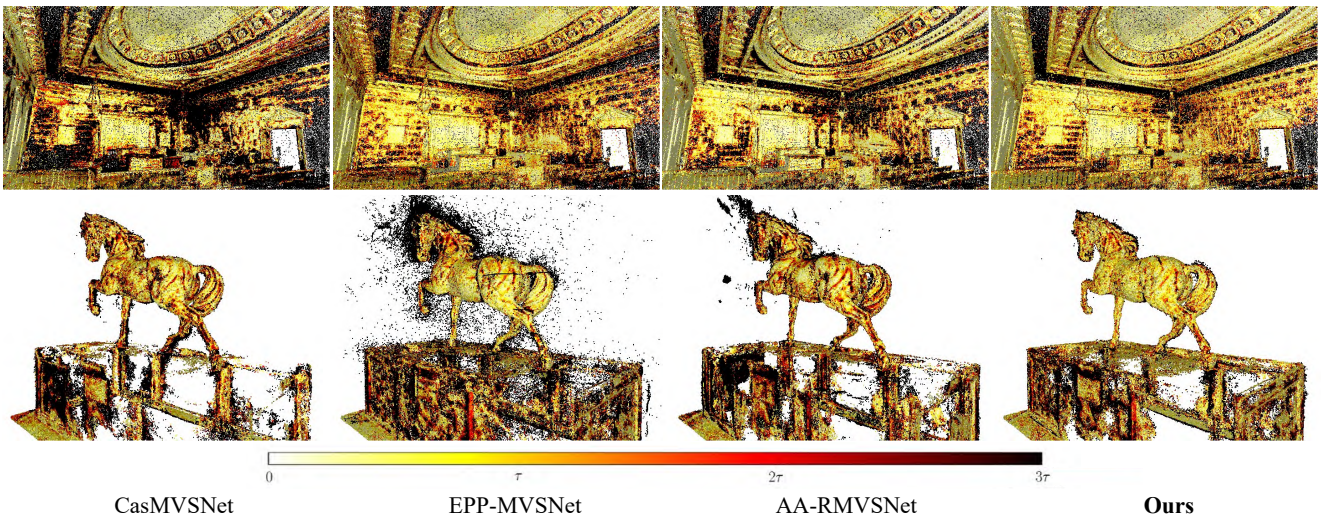

Figure 6. Comparison of reconstructed results with several state-of-the-art methods [3, 10, 27] on Tanks and Temples benchmark [12]. τ is the scene-relevant distance threshold determined officially and darker regions indicate larger error encountered with regard to τ. The first row shows Recall on the scene of Courtroom ($\tau = 10mm$); the second row shows Precision on the scene of Horse ($\tau = 3mm$).

metrics described in [7].

Some quantitative results are illustrated in Tab. 3. EPE stands for the endpoint error, which is the average ℓ-1 distance between the prediction and the ground truth depth; e_1 and e_3 represent the proportion in % of pixels with depth error larger than 1 and larger than 3. Compared with other methods, TransMVSNet achieves impressive results, demonstrating its capability of yielding high-quality depth maps. Please refer to the Supplementary Material for more point cloud results.

4.4. Ablation Study

We perform ablation studies to analyze the effectiveness and costs of different modules. The implemented baseline is basically based on CasMVSNet [10], which applies feature correlation and is trained with ℓ-1 loss. All the experiments are performed with the same hyperparameters.

Method	EPE	e_1	e_3
MVSNet [31]	1.49	21.98	8.32
CVP-MVSNet [30]	1.90	19.73	10.24
CasMVSNet [10]	1.43	19.01	9.77
Vis-MVSNet [34]	1.47	15.14	5.13
EPP-MVSNet [19]	1.17	12.66	6.20
TransMVSNet	**0.73**	**8.32**	**3.62**

Table 3. Quantitative results towards predicted depth maps on BlendedMVS validation set [33] (**lower is better**).

As shown in Tab. 4, after applying focal loss, the overall performance improves by 1.7% while the computational costs remain unchanged. Due to the computational efficiency of Linear Transformer, we are able to leverage FMT with little additional costs in terms of memory and MACs but its inference speed is nearly 1.4 times slower. With the transformed feature pathway, both Completeness and Overall performance get boosted while there is almost no

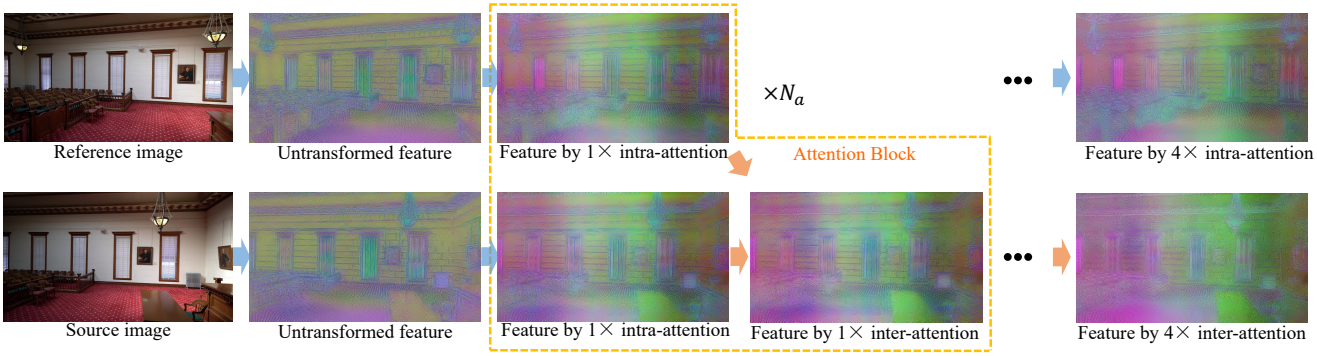

Figure 7. Evolution of feature maps via FMT on the scene Courtroom of Tanks and Temples benchmark [12]. We apply PCA to reduce the number of feature channels to 3 and visualize the results with RGB. Images of the first row show features of the reference view, which are only updated by intra-attention in FMT; images of the second row represent features of a source view, which get updated by both intra- and inter-attention layers.

	Model Settings				Mean Distance			Mem.	MACs	Time
	F.L.	FMT	Pathway	ARF	Acc.	Comp.	Overall			
(a)					0.351	0.339	0.345	3244	212	0.271
(b)	✓				0.343	0.335	0.339	3244	212	0.271
(c)	✓	✓			0.335	0.310	0.323	3288	235	0.638
(d)	✓	✓	✓		0.332	0.298	0.315	3288	241	0.677
(e)	✓	✓	✓	✓	**0.321**	**0.289**	**0.305**	3778	435	0.996

Table 4. Quantitative performance with different components on DTU evaluation dataset [1]. F.L. is short for focal loss. The unit is *MB* for memory occupancy (Mem.), *G* for multiply–accumulate operations (MACs) and *second* for inference time.

increase in its memory occupancy, indicating the effectiveness and efficiency of the pathway. With ARF module attached, the full TransMVSNet is able to achieve state-of-the-art performance by a large margin. ARF module brings considerable computational costs in all aspects. After all, the inference time is still within one second, which is acceptable compared to RNN-based methods [27, 29, 32].

5. Discussions

5.1. Comparisons to Related Work

TransMVSNet vs. CasMVSNet Our architecture is based on the coarse-to-fine regularization pattern proposed by CasMVSNet [10]. The main difference is that we introduce Transformer to capture long-range global context for better feature matching over multiple views. Using the coarse-to-fine manner brings more computation efficiency while remarkable performance is also achieved.

TransMVSNet vs. LoFTR LoFTR [24] interleaves self- and cross-attention layers multiple times along flattened feature maps to estimate dense matching between a pair of images. Different from one-to-one matching tasks, MVS is actually a one-to-many matching task. We thus propose the FMT module to adapt attention layers to MVS.

TransMVSNet vs. STTR STTR [14] performs self- and cross-attention along intra- and inter-epipolar line to estimate stereo depth, where the context range of local features is only limited to their corresponding epipolar lines. Note that there does not exist line-to-line correspondence in MVS, and we thus utilize attention layers along whole flattened feature maps, to bring global context into feature matching over multiple views.

5.2. Limitations

- Transformer slows down the speed of inference, as is shown in Tab. 4.
- Similar to other coarse-to-fine MVS networks, our method is sensitive to inference hyperparameters, *e.g.* number of depth hypotheses, depth interval, and decay factor of depth interval.

6. Conclusion

In this paper, we present a novel learning-based MVS network, termed as TransMVSNet, which aggregates global long-range context-aware information via Transformer. Specifically, TransMVSNet comprises an effective Feature Matching Transformer (FMT) module formulated with intra-attention and inter-attention, which focus on retrieving context-aware information within and across images respectively. Moreover, we design the Adapative Receptive Field (ARF) module and a transformed feature pathway to better facilitate the function of FMT. Extensive experiments show that TransMVSNet achieves state-of-the-art performance on DTU dataset, Tanks and Temples benchmark, and Blended-MVS dataset.

Acknowledgement This paper is supported by the National Key R&D Plan of the Ministry of Science and Technology (Project No. 2020AAA0104400).

References

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120(2):153–168, 2016. 1, 2, 5, 6, 8

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 2

[3] Shuo Cheng, Zexiang Xu, Shilin Zhu, Zhuwen Li, Li Erran Li, Ravi Ramamoorthi, and Hao Su. Deep stereo using adaptive thin volume representation with uncertainty awareness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2524–2534, 2020. 1, 2, 6, 7

[4] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). In *International Conference on Learning Representations*, 2016. 4

[5] Robert T Collins. A space-sweep approach to true multi-image matching. In *Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 358–363. IEEE, 1996. 1

[6] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 5

[7] François Darmon, Bénédicte Bascle, Jean-Clément Devaux, Pascal Monasse, and Mathieu Aubry. Deep multi-view stereo gone wild. *arXiv preprint arXiv:2104.15119*, 2021. 7

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 2

[9] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 873–881, 2015. 6

[10] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping 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*, pages 2495–2504, 2020. 1, 2, 5, 6, 7, 8

[11] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, pages 5156–5165. PMLR, 2020. 4

[12] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017. 1, 2, 5, 6, 7, 8

[13] Andreas Kuhn, Christian Sormann, Mattia Rossi, Oliver Erdler, and Friedrich Fraundorfer. Deepc-mvs: Deep confidence prediction for multi-view stereo reconstruction. In *2020 International Conference on 3D Vision (3DV)*, pages 404–413. IEEE, 2020. 7

[14] Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X Creighton, Russell H Taylor, and Mathias Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6197–6206, 2021. 2, 8

[15] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 3, 4

[16] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 2, 5

[17] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021. 2

[18] Keyang Luo, Tao Guan, Lili Ju, Yuesong Wang, Zhuo Chen, and Yawei Luo. Attention-aware multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1590–1599, 2020. 7

[19] Xinjun Ma, Yue Gong, Qirui Wang, Jingwei Huang, Lei Chen, and Fan Yu. Epp-mvsnet: Epipolar-assembling based depth prediction for multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5732–5740, 2021. 1, 6, 7

[20] Xuran Pan, Zhuofan Xia, Shiji Song, Li Erran Li, and Gao Huang. 3d object detection with pointformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7463–7472, 2021. 2

[21] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021. 2

[22] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 1, 2, 4

[23] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision*, pages 501–518. Springer, 2016. 6, 7

[24] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8922–8931, 2021. 1, 2, 8

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural*

information processing systems, pages 5998–6008, 2017. 1, 2, 4

[26] Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Pablo Speciale, and Marc Pollefeys. Patchmatchnet: Learned multi-view patchmatch stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14194–14203, 2021. 1, 6, 7

[27] Zizhuang Wei, Qingtian Zhu, Chen Min, Yisong Chen, and Guoping Wang. Aa-rmvsnet: Adaptive aggregation recurrent multi-view stereo network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6187–6196, 2021. 1, 2, 6, 7, 8

[28] Qingshan Xu and Wenbing Tao. Multi-scale geometric consistency guided multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5483–5492, 2019. 7

[29] Jianfeng Yan, Zizhuang Wei, Hongwei Yi, Mingyu Ding, Runze Zhang, Yisong Chen, Guoping Wang, and Yu-Wing Tai. Dense hybrid recurrent multi-view stereo net with dynamic consistency checking. In *European Conference on Computer Vision*, pages 674–689. Springer, 2020. 2, 6, 8

[30] Jiayu Yang, Wei Mao, Jose M Alvarez, and Miaomiao Liu. Cost volume pyramid based depth inference for multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4877–4886, 2020. 2, 5, 7

[31] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 767–783, 2018. 1, 2, 5, 7

[32] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5525–5534, 2019. 1, 2, 6, 7, 8

[33] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long 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*, pages 1790–1799, 2020. 5, 6, 7

[34] Jingyang Zhang, Yao Yao, Shiwei Li, Zixin Luo, and Tian Fang. Visibility-aware multi-view stereo network. *British Machine Vision Conference (BMVC)*, 2020. 1, 2, 5, 6, 7

[35] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9308–9316, 2019. 5