# Supplementary Material for TransMVSNet: Global Context-aware Multi-view Stereo Network with Transformers

## 1. About One-to-many Matching Pattern

As has been mentioned in the main paper, the nature of multi-view stereo (MVS) is a one-to-many matching task. We here further discuss the rationality of this analogy. As is illustrated in Fig. 1 where we consider a simple case that $N = 3$, for a pixel in the reference image, we attempt to find an optimal depth value $d$ among plane sweep depth hypotheses. These candidate 3D points all lie in an epipolar line of neighboring source images. In this way, the task of MVS becomes a typical one-to-many matching task where each pixel in $\mathbf{I}_0$ is supposed to find the best match among candidate source points in $\mathbf{I}_1$, and also, among candidate source points in $\mathbf{I}_2$.

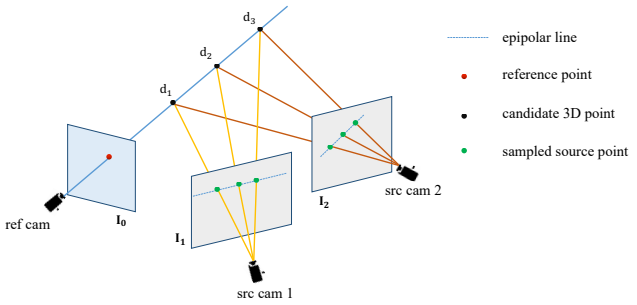



Figure 1. Illustration of the matching pattern of MVS.

## 2. Ablation Study on Hyperparameters

### 2.1. Number of Views & Input Resolution

We conduct ablation study against the number of input views $N$ and input resolution $H \times W$ on DTU evaluation set [1], and the results are listed in Tab. 1.

| $N$ | $H \times W$ | Acc. | Comp. | Overall |
|---|---|---|---|---|
| 3 | $864 \times 1152$ | **0.310** | 0.323 | 0.316 |
| 5 | $864 \times 1152$ | 0.321 | **0.289** | **0.305** |
| 7 | $864 \times 1152$ | 0.327 | 0.295 | 0.311 |
| 9 | $864 \times 1152$ | 0.327 | 0.308 | 0.317 |
| 5 | $512 \times 640$ | 0.405 | 0.319 | 0.361 |

Table 1. Ablation study on number of input views $N$ and image resolution $H \times W$ on DTU evaluation set [1] (**lower is better**).

### 2.2. Focusing Parameter

We study the influence of training with different focusing parameter $\gamma$ of focal loss [4] in Tab. 2. Other inference settings are the same as the reported case in the main paper. By experiment results, $\gamma = 0$ best fits the scene complexity of DTU dataset [1], and obtains best Accuracy and Overall scores. As for BlendedMVS dataset [8], whose scenes are more diverse and complicated, $\gamma = 2$ makes a big difference.

| $\gamma$ | DTU | | | BlendedMVS | | |
|---|---|---|---|---|---|---|
| | Acc. | Comp. | Overall | EPE | $e_1$ | $e_3$ |
| 0 | **0.321** | 0.289 | **0.305** | 0.80 | 9.79 | 4.40 |
| 0.5 | 0.341 | 0.282 | 0.312 | 0.79 | 8.91 | 3.94 |
| 1 | 0.342 | **0.279** | 0.310 | 0.78 | 8.80 | 3.86 |
| 2 | 0.345 | 0.282 | 0.314 | **0.73** | **8.32** | **3.62** |

Table 2. Ablation study on the value of $\gamma$ on DTU evaluation set [1] and BlendedMVS validation set [8] (**lower is better**).

## 3. Ablation Study on FMT Design

We further explore the architecture design of the Feature Matching Transformer (FMT) described in the main paper.

### 3.1. Number of Attention Heads

As is mentioned in the main paper, we apply multi-head attention where feature channels are split into $N_h$ groups, namely attention heads. We therefore study the influence of $N_h$, including model performance, memory consumption and inference time, in Tab. 3. With $N_h$ increasing, there is no difference in memory consumption, but the inference time fluctuates possibly due to PyTorch's underlying implementation.

| $N_h$ | Acc. | Comp. | Overall | Mem.(*MB*) | Time(*s*) |
|---|---|---|---|---|---|
| 1 | 0.322 | 0.296 | 0.309 | 3778 | **0.706** |
| 2 | 0.322 | 0.290 | 0.306 | 3778 | 1.024 |
| 4 | 0.322 | 0.289 | 0.306 | 3778 | 1.020 |
| 8 | **0.321** | **0.289** | **0.305** | 3778 | 0.996 |
| 16 | **0.321** | 0.290 | 0.306 | 3778 | 1.040 |

Table 3. Ablation study on the number of attention heads $N_h$ on DTU evaluation set [1] (**lower is better**).

### 3.2. Number of Attention Blocks

We adjust the number of attention blocks $N_a$ and Tab. 4 shows respective evaluation results, memory occupancy and inference time. As is demonstrated, $N_a = 4$ achieves a balance between performance and efficiency.

| $N_a$ | Acc. | Comp. | Overall | Mem.(*MB*) | Time(*s*) |
|---|---|---|---|---|---|
| 2 | 0.337 | **0.288** | 0.312 | **3760** | **0.815** |
| 4 | 0.321 | 0.289 | **0.305** | 3778 | 0.996 |
| 6 | **0.319** | 0.298 | 0.308 | 3792 | 1.223 |

Table 4. Ablation study on the number of attention blocks on DTU evaluation set [1] (**lower is better**).

### 3.3. Design of Attention Block

During the exploration, we study several potential architectures for attention block. We assume that intra-attention is always performed upon both the reference feature $\mathcal{F}_0$ and source features $\{\mathcal{F}_i\}_{i=1}^{N-1}$. Therefore the main differences lie in how to handle the reference feature $\mathcal{F}_0$ under inter-attention. We present all 4 possible designs covered as follows. The 4 possible candidate designs are illustrated in Fig. 2.

(a) Only reference-to-source inter-attention is performed upon $\mathcal{F}_0$, so $\mathcal{F}_0$ is only updated by intra-attention. **This is the final choice in TransMVSNet.**

(b) We sort $N-1$ source images with the same view selection protocol used in [6] and perform source-to-reference inter-attention sequentially. Then the reference-to-source inter-attention is done.

(c) We duplicate $\mathcal{F}_0$ into $N-1$ identical copies so that source-to-reference inter-attention can be performed pairwise in parallel. For the reference-to-source inter-attention, we average all $N-1$ transformed $\mathcal{F}_0$.

(d) $\mathcal{F}_0$ is duplicated at the very beginning of FMT so that each source and the reference feature form a pair throughout the whole FMT. Inter-attention operations of both directions, in this way, are performed at the same time and the final outputs of FMT are $N-1$ respectively transformed $\mathcal{F}_{0,i}(i = 1, \ldots, N-1)$ and $N-1$ transformed source features.

To conclude, $\mathcal{F}_0$ in (b)(c)(d) is updated by source-to-reference inter-attention. (b) and (c) differ in the order of inter-attention: (b) does it sequentially while (c) is in parallel. (a)(b)(c) all follow a one-to-many matching pattern while (d) explicitly splits it into $N-1$ one-to-one problems. We also quantitatively study different designs in terms of both performance and costs. As is shown in Tab. 5, memory consumption of (a)(b)(c) is identical but (a) is more efficient in inference time. (d) occupies significantly more memory and takes more time than other candidates. As a result, (a) is both effective and efficient, verifying the intuition that under a one-to-many matching scenario, the matching target (the reference feature) should always be identical.

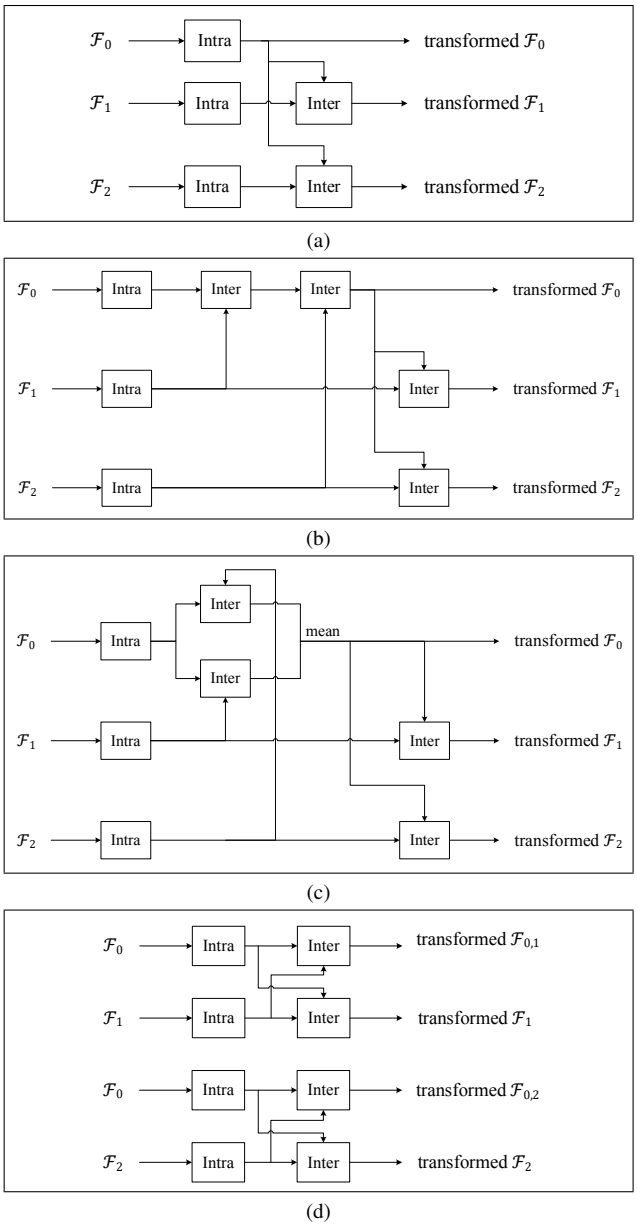



Figure 2. Candidate designs of attention block. Note that there are in total $N-1$ source feature maps and we omit most of them for brevity.

| | Matching Pattern | $\mathcal{F}_0$ Updated by Inter-att. | Acc. | Comp. | Overall | Mem. (*MB*) | Time (*s*) |
|---|---|---|---|---|---|---|---|
| (a) | one-to-many | no | 0.321 | **0.289** | **0.305** | **3778** | **0.996** |
| (b) | one-to-many | yes | **0.320** | 0.304 | 0.312 | **3778** | 1.197 |
| (c) | one-to-many | yes | 0.332 | 0.294 | 0.313 | **3778** | 1.178 |
| (d) | one-to-one | yes | 0.339 | 0.292 | 0.316 | 4142 | 1.331 |

Table 5. Ablation study on the architecture design of attention blocks on DTU evaluation set [1] (**lower is better**).

## 4. Visualized Attention

We visualize the weights of both intra- and inter-attention in Fig. 3. For query points from challenging regions, *e.g*. textureless or non-Lambertian surfaces, intra-attention seeks context information globally and inter-attention tends to match features across images. These two attention mechanisms are complementary and beneficial to robust depth estimation at challenging regions.

## 5. Visualized Feature Map

To better illustrate the evolution of feature maps throughout FMT, we give another group of examples to demonstrate how FMT changes the feature in Fig. 4. Before FMT, the extracted feature maps from FPN are not recognizable enough for robust feature matching at challenging areas since the feature representation is mostly local. After several intra- and inter-attention modules, more global position-dependent context information is encoded into the feature map, which benefits feature matching for textureless and non-Lambertian surfaces.

## 6. More Point Cloud Results

We visualize all results of DTU evaluation set [1], the intermediate and advanced set of Tanks and Temples benchmark [3] and BlendedMVS validation set [8] respectively in Fig. 5, Fig. 6 and Fig. 7. Our TransMVSNet demonstrates its robustness and scalability on scenes with varying depth ranges.

## 7. Use of Existing Assets

The implementation of TransMVSNet is based on Cas-MVSNet [2], who also heavily borrows code from the PyTorch version of MVSNet [6].

Preprocessed images and camera parameters of both DTU dataset [1] and Tanks and Temples benchmark [3] are from the official repository of MVSNet [6] & R-MVSNet [7], where COLMAP-SfM [5] is adopted to obtain the camera calibration for Tanks and Temples.

## 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, 3, 6

[2] 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. 3

[3] 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. 3, 7

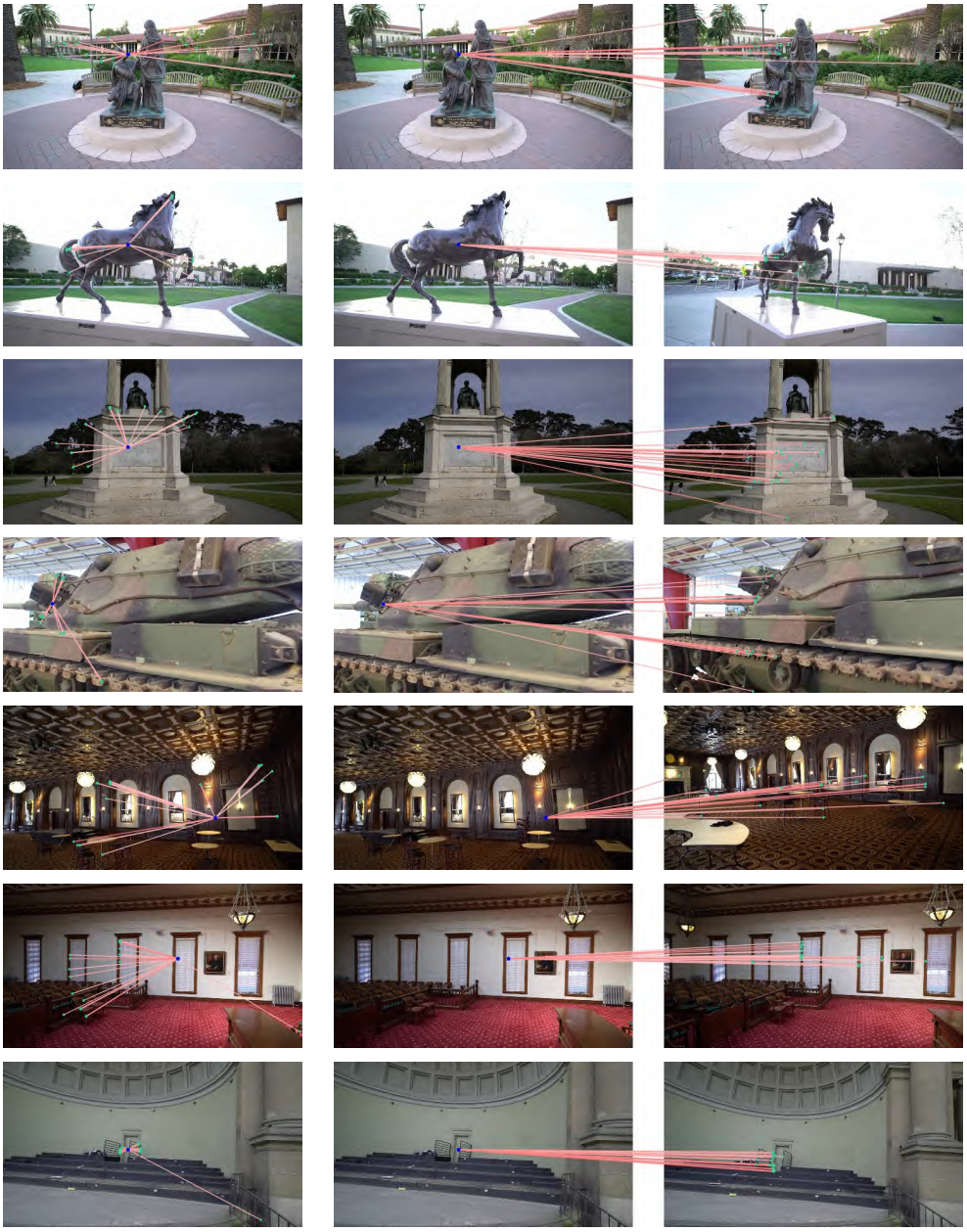
[4] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 1

[5] Johannes L Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 3

[6] 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. 2, 3

[7] 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. 3

[8] 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. 1, 3, 8



(a) Source view intra-attention (b) Reference-to-source inter-attention

Figure 3. Visualization of intra- and inter-attention weights. From left to right, the first column (a) shows the intra-attention weights of a query point in the source image; the second and the third columns (b) show the inter-attention weights of the same query point in the source image with regard to its top-20 correspondences in the reference image.

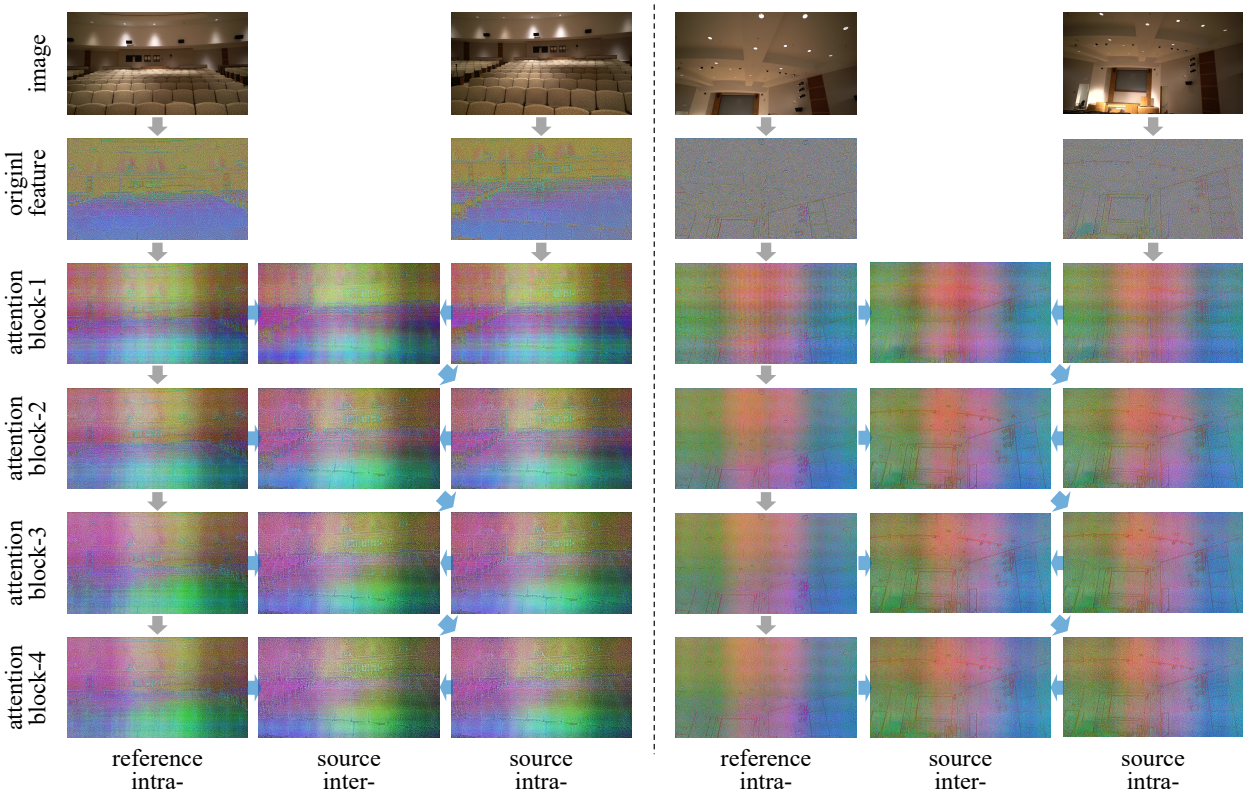



Figure 4. Evolution of transformed feature after each attention block. For better visualization, we apply PCA to reduce the number of feature channels to 3 and color the channels with RGB.


Figure 5. Point clouds of all 22 scans in DTU evaluation set [1] reconstructed by TransMVSNet.


Figure 6. All point clouds of Tanks and Temples Benchmark [3] (intermediate & advanced) reconstructed by TransMVSNet.

Figure 7. Point clouds of all 7 scenes in BlendedMVS validation set [8] reconstructed by TransMVSNet.