Supplementary Material for MVSTER: Epipolar Transformer for Efficient Multi-View Stereo

1 Discussion

Analysis of attention formulation The attention module is designed to enhance the fusion of the cost volumes from different views. Therefore, it is straightforward to use view-wise attention (VA) as a guidance for fusion. However, VA ignores depth-wise associations, which is critical for reducing depth ambiguity in cost volumes. Therefore, the proposed epipolar Transformer utilizes depth-wise attention (DA) to fuse cost volumes, which highlights the matched depth bin and surpasses irrelevant ones. To demonstrate the effectiveness of DA, we compare DA with VA on DTU. Specifically, the overall score of DA (0.313) relatively improves the baseline (0.323) by 3.1%, which validates that the relation across depth hypotheses is more beneficial for depth prediction.

The novelty compared with MVS2D [8] (*i*) MVS2D uses epipolar attention module to implicitly embed depth information into 2D feature maps. In contrast, our epipolar Transformer leverages depth-wise attention to fuse 3D cost volumes, which explicitly maintains the importance of each depth hypothesis and makes depth prediction more tractable. (*ii*) MVS2D introduces additional trainable linear maps to transform features before cross attention, while the proposed epipolar Transformer learns data-dependent associations without introducing learnable parameters. We further compare model performance and efficiency on DTU. Specifically, MVSTER (0.303@0.17s) shows comparable efficiency with MVS2D (0.342@0.13s). Besides, MVSTER relatively improves the overall score by 11.4%, which demonstrates that explicit usage of 3D depth-wise attention is beneficial for depth prediction.

2 Additional Implementation Details

Network Architecture of Feature Extractor We use a four-stage Feature Pyramid Network (FPN) [5] to extract image features, and the detailed parameters with layer descriptions are summarized in Table 1. For Deformable Convolutional Networks (DCN) [3] and Atrous Spatial Pyramid Pooling (ASPP) [2] that are used in the ablation study of our main text, the network parameters are listed in Table 2.

Table 1. The detailed parameters of FPN, where S denotes stride, and if not specified with *, each convolution layer is followed by a Batch Normalization layer (BN) and a Rectified Linear Unit (ReLU).

Stage Description	Layer Description	Output Size
-	Input Images	$H \times W \times 3$
FPN Stage 1	Conv2D, 3×3, S1, 8	$H \times W \times 8$
FPN Stage 1	Conv2D, 3×3, S1, 8	$H \times W \times 8$
FPN Stage 1 Inner Layer*	Conv2D, 1×1, S1, 64	$H \times W \times 64$
FPN Stage 1 Output Layer*	Conv2D, 1×1, S1, 8	$H \times W \times 8$
FPN Stage 2	Conv2D, 5×5, S2, 16	$H/2 \times W/2 \times 16$
FPN Stage 2	Conv2D, 3×3, S1, 16	$H/2 \times W/2 \times 16$
FPN Stage 2	Conv2D, 3×3, S1, 16	$H/2 \times W/2 \times 16$
FPN Stage 2 Inner Layer*	Conv2D, 1×1, S1, 64	$H/2 \times W/2 \times 64$
FPN Stage 2 Output Layer*	Conv2D, 1×1, S1, 16	$H/2 \times W/2 \times 16$
FPN Stage 3	Conv2D, 5×5, S2, 32	$H/4 \times W/4 \times 32$
FPN Stage 3	Conv2D, 3×3, S1, 32	$H/4 \times W/4 \times 32$
FPN Stage 3	Conv2D, 3×3, S1, 32	$H/4 \times W/4 \times 32$
FPN Stage 3 Inner Layer*	Conv2D, 1×1, S1, 64	$H/4 \times W/4 \times 64$
FPN Stage 3 Output Layer*	Conv2D, 1×1, S1, 32	$H/4 \times W/4 \times 32$
FPN Stage 4	Conv2D, 5×5, S2, 64	$H/8 \times W/8 \times 64$
FPN Stage 4	Conv2D, 3×3, S1, 64	$H/8 \times W/8 \times 64$
FPN Stage 4	Conv2D, 3×3, S1, 64	$H/8 \times W/8 \times 64$
FPN Stage 4 Inner Layer*	Conv2D, 1×1, S1, 64	$H/8 \times W/8 \times 64$
FPN Stage 4 Output Layer*	Conv2D, 1×1, S1, 64	$H/8 \times W/8 \times 64$

Network Architecture of Light-Weight 3D CNN An UNet [6] structured 3D CNN is applied for cost volume regularization at each stage, where the kernel size $3 \times 3 \times 3$ is partially replaced with $3 \times 3 \times 1$ in MVSTER for a more efficient pipeline. Apart from the input cost volume size, the network architectures are the same for each stage in the cascade structure, so we only report the detailed parameters of the 4th stage in Table 3.

3 Additional Ablation Study

Ablation Study on Hyperparameters We conduct an ablation study on loss weight λ and temperature parameter t_e. As shown in Table 4, $\lambda = 3 \times 10^{-4}$ is a proper loss weight for jointly optimizing monocular depth estimation and multi-view stereo. As shown in Table 5, MVSTER produces a finer depth map when slowly increasing t_e, and the network shows best reconstruction performance on DTU when $t_e = 2$.

Table 2. The detailed parameters of DCN and ASPP, where S denotes stride, and D denotes dilation parameter for ASPP.

Stage Description	Layer Description	Output Size
DCN Stage 1	DCN2D, 3×3, S1, 8	$H \times W \times 8$
DCN Stage 2	DCN2D, 3×3, S1, 16	$H/2 \times W/2 \times 16$
DCN Stage 3	DCN2D, 3×3, S1, 32	$H/4 \times W/2 \times 32$
DCN Stage 4	DCN2D, 3×3, S1, 64	$H/8 \times W/8 \times 64$
ASPP Stage 1	Conv2D, 3×3, S1, D{1,6,12}, 8	$H \times W \times 8$
ASPP Stage 2	Conv2D, 3×3, S1, D{1,6,12}, 16	$H/2 \times W/2 \times 16$
ASPP Stage 3	Conv2D, 3×3, S1, D{1,6,12}, 32	$H/4 \times W/2 \times 32$
ASPP Stage 4	Conv2D, 3×3, S1, D{1,6,12}, 64	$H/8 \times W/8 \times 64$

Table 3. The detailed parameters of 3D CNN, where S denotes stride, and if not specified with *, each convolution layer is followed by a Batch Normalization layer (BN) and a Rectified Linear Unit (ReLU).

Stage Description	Layer Description	Output Size
-	Input Cost Volume	$H \times W \times 4 \times 8$
UNet Stage 1	Conv3D, $3 \times 3 \times 1$, S1, 8	$H \times W \times 4 \times 8$
UNet Stage 1	Conv3D, $3 \times 3 \times 1$, S2, 16	$H/2 \times W/2 \times 4 \times 16$
UNet Stage 1	Conv3D, $3 \times 3 \times 3$, S1, 16	$H/2 \times W/2 \times 4 \times 16$
UNet Stage 1 Inner Layer	TransposeConv3D, $3 \times 3 \times 1$, S2, 8	$H \times W \times 4 \times 8$
UNet Stage 1 Output Layer*	TransposeConv3D, $3 \times 3 \times 3$, S1, 8	$H \times W \times 4 \times 8$
UNet Stage 2	Conv3D, $3 \times 3 \times 1$, S2, 32	$H/4 \times W/4 \times 4 \times 32$
UNet Stage 2	Conv3D, $3 \times 3 \times 3$, S1, 32	$H/4 \times W/4 \times 4 \times 32$
UNet Stage 2 Inner Layer	TransposeConv3D, $3 \times 3 \times 1$, S2, 16	$H/2 \times W/2 \times 4 \times 16$
UNet Stage 3	Conv3D, $3 \times 3 \times 1$, S2, 64	$H/8 \times W/8 \times 4 \times 64$
UNet Stage 3	Conv3D, $3 \times 3 \times 3$, S1, 64	$H/8 \times W/8 \times 4 \times 64$
UNet Stage 3 Inner Layer	TransposeConv3D, $3 \times 3 \times 1$, S2, 32	$H/4 \times W/4 \times 4 \times 32$

4 Point Cloud Visualizations

We visualize point cloud reconstruction results of DTU [1], ETH3D [7] and Tanks&Temples [4] in Fig. 1, Fig. 2 and Fig. 3, respectively. MVSTER shows its robustness on scenes with varying input image resolutions and depth ranges.

Fig. 1. Point clouds on DTU [1] reconstructed by MVSTER.

Fig. 2. Point clouds on ETH3D [7] reconstructed by MVSTER.

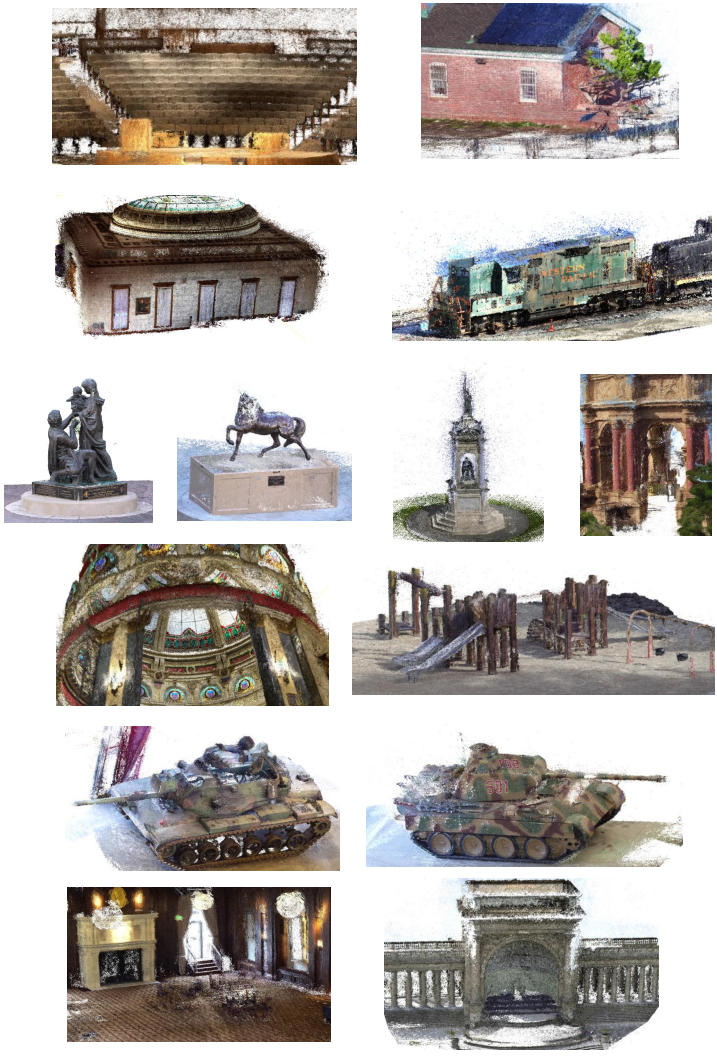

Fig. 3. Point clouds on Tanks&Temples [4] reconstructed by MVSTER.

Table 4. Ablation study on loss weight λ.

λ	Acc.↓	Comp.↓	Overall↓	EPE↓	e_1 ↓	e_3 ↓
1×10^{-2}	0.361	0.312	0.337	1.56	16.47	8.32
1×10^{-3}	0.354	0.287	0.321	1.33	15.01	**7.26**
3×10^{-4}	**0.350**	0.276	**0.313**	**1.31**	14.98	7.27
1×10^{-4}	0.354	**0.275**	0.314	1.32	**14.97**	7.28

Table 5. Ablation study on temperature parameter t_e.

λ	Acc.↓	Comp.↓	Overall↓	EPE↓	e_1 ↓	e_3 ↓
0.5	0.354	0.279	0.317	1.33	15.09	7.52
1.0	0.353	0.279	0.314	1.33	15.03	7.47
2.0	**0.350**	0.276	**0.313**	**1.31**	14.98	7.27
3.0	0.353	**0.274**	0.314	**1.31**	**14.89**	**7.08**

References

1. Aanæs, H., Jensen, R.R., Vogiatzis, G., Tola, E., Dahl, A.B.: Large-scale data for multiple-view stereopsis. International Journal of Computer Vision (2016)
2. Chen, L., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
3. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: IEEE International Conference on Computer Vision (2017)
4. Knapitsch, A., Park, J., Zhou, Q., Koltun, V.: Tanks and temples: benchmarking large-scale scene reconstruction. ACM Transactions on Graphics (2017)
5. Lin, T., Dollár, P., Girshick, R.B., He, K., Hariharan, B., Belongie, S.J.: Feature pyramid networks for object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (2017)
6. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention (2015)
7. Schöps, T., Schönberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: IEEE Conference on Computer Vision and Pattern Recognition (2017)
8. Yang, Z., Ren, Z., Shan, Q., Huang, Q.: MVS2D: efficient multi-view stereo via attention-driven 2d convolutions. arXiv preprint arXiv:2104.13325 (2021)