MonoMVSNet: Monocular Priors Guided Multi-View Stereo Network

Supplementary Material

A. Details of Camera Embedding

The details of camera embedding module are as follows. The warped feature ($B \times C \times D \times H \times W$) is reshaped to $B \times CD \times H \times W$ and further processed by Conv→BN→ReLU layers to produce the output with shape $B \times C \times H \times W$. The camera parameters ($B \times 4 \times 4$) is reshaped to $B \times 16$, processed by a BN layer, and mapped to $B \times C$ by an MLP, which is further reshaped to $B \times C \times 1 \times 1$. The two features ($B \times C \times H \times W$ and $B \times C \times 1 \times 1$) are added together with brodcast, wich is further processed by Squeeze-and-Excitation [4]→Conv layers. The output is $B \times C \times H \times W$, which then add to the corresponding feature.

B. More Ablation Study

Compatibility with Monocular Foundation Models. To fairly verify the applicability of our method to different monocular foundation models, we replace Depth Anything V2 [10] with other ViT-small version monocular foundation models: DINO V2 [7], Depth Anything V1 [9], and Depth Pro [2]. As shown in Tab. 1, all alternative monocular foundation models exhibit significant improvements in both point cloud and depth performance, among which Depth Anything V2 achieves the best results. This further confirms the generalization capability of our method.

Models	Overall↓	Acc.↓	Comp.↓	MAE↓
Depth Pro [2]	0.286	0.316	0.256	5.78
DINO V2 [7]	0.284	0.311	0.257	5.50
Depth Anything V1 [9]	0.282	0.299	0.265	5.03
Depth Anything V2 [10]	0.278	0.313	0.243	4.99

Table 1. Ablation study on different monocular foundation models.

Number of Views. As shown in Tab. 2, we show the impact of the number of input views. Multi-view information helps alleviate problems such as occlusions, and the reconstruction quality progressively improves with an increasing number of views, saturating at 9 views.

Positional Encoding Design. We replace Cross-View Positional Encoding (CVPE) with an MLP to map the traditional 2D positional encoding. As shown in Fig. 3, the Overall↓ metric (0.278→0.285), demonstrating the effectiveness our proposed CVPE.

C. More Visualization Results

Fig. 1 present our method achieves better accuracy and recall in textureless regions and depth discontinuous edge regions on the Tanks-and-Temples benchmark [5]. Figure 2 presents the depth map comparison results from the ablation experiments on the DTU [1] dataset. Figures 3 and 4 visualize the reconstructed point clouds on the DTU and Tanks-and-Temples benchmark, respectively.

N	Overall↓	Acc.↓	Comp.↓
4	0.2825	0.315	0.250
5	0.2780	0.313	**0.243**
6	0.2765	0.309	0.244
7	0.2760	0.307	0.245
8	0.2755	0.304	0.247
9	**0.2750**	0.302	0.248
10	0.2755	**0.299**	0.252

Table 2. Ablation study on the number of input views N.

Position Encoding	Overall↓	MAE↓
w/ 2D PE+MLP	0.285	5.29
w/ 2D PE+CVPE	0.278	4.99

Table 3. Ablation on position encoding design.

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*, 120:153–168, 2016. 1, 3, 4

[2] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024. 1

[3] Chenjie Cao, Xinlin Ren, and Yanwei Fu. Mvsformer++: Revealing the devil in transformer's details for multi-view stereo. *arXiv preprint arXiv:2401.11673*, 2024. 2

[4] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 1

[5] 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] Tianqi Liu, Xinyi Ye, Weiyue Zhao, Zhiyu Pan, Min Shi, and Zhiguo Cao. When epipolar constraint meets non-local operators in multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18088–18097, 2023. 2

[7] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy

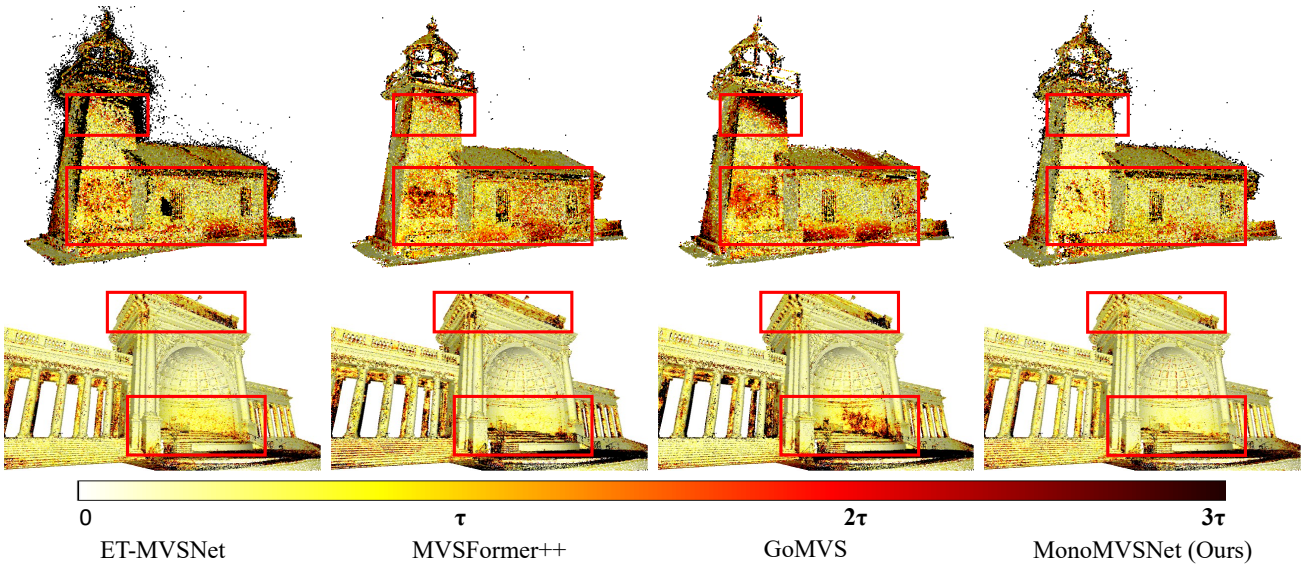

Figure 1. Qualitative comparison of reconstructed point clounds with ET-MVSNet [6], MVSFomrer++ [3], and GoMVS [8] on Tanks-and-Temples [5] benchmark. Brighter areas in the figure indicate smaller errors associated with the distance threshold (τ). The top row shows the Precision for the Lighthouse in the advanced subset ($\tau = 5mm$), the bottom row shows the Recall for the Temple in the intermediate subset ($\tau = 15mm$).

Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1

[8] Jiang Wu, Rui Li, Haofei Xu, Wenxun Zhao, Yu Zhu, Jinqiu Sun, and Yanning Zhang. Gomvs: Geometrically consistent cost aggregation for multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20207–20216, 2024. 2

[9] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 1

[10] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. 1

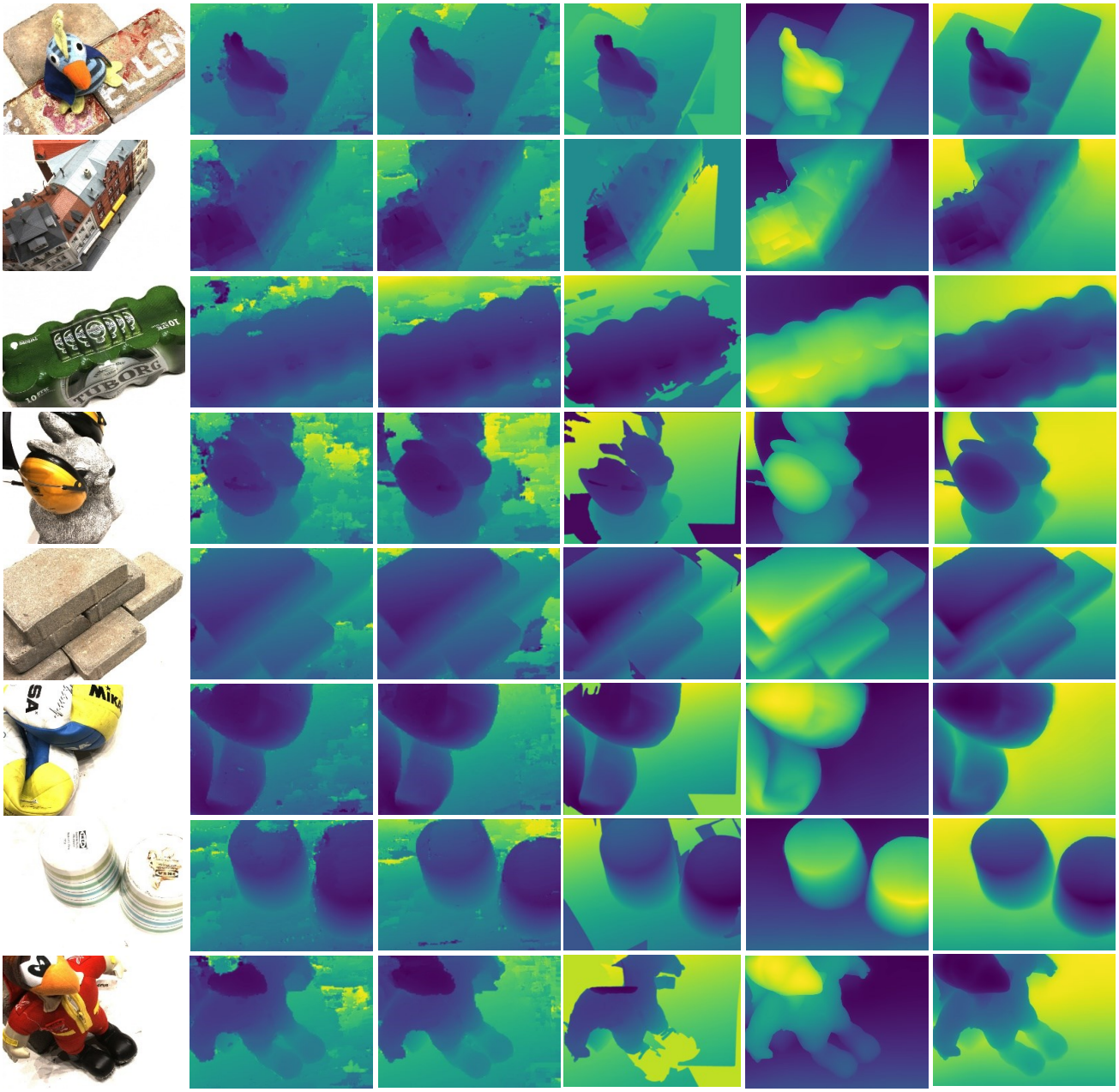

Figure 2. Additional depth maps visualization comparing the Baseline, MonoMVSNet, Ground Truth, Depth Anything V2 (DAV2), and Aligned DAV2 on the DTU [1] dataset.

Figure 3. More visualization results of all reconstructed point clouds on the DTU [1] dataset.

Figure 4. More visualization results of all reconstructed point clouds on the Tanks-and-Temples [5] benchmark.