SPE-MVS: Spatial Position Encoding Enhanced Multi-View Stereo with Monocular Depth Priors

Supplementary Material

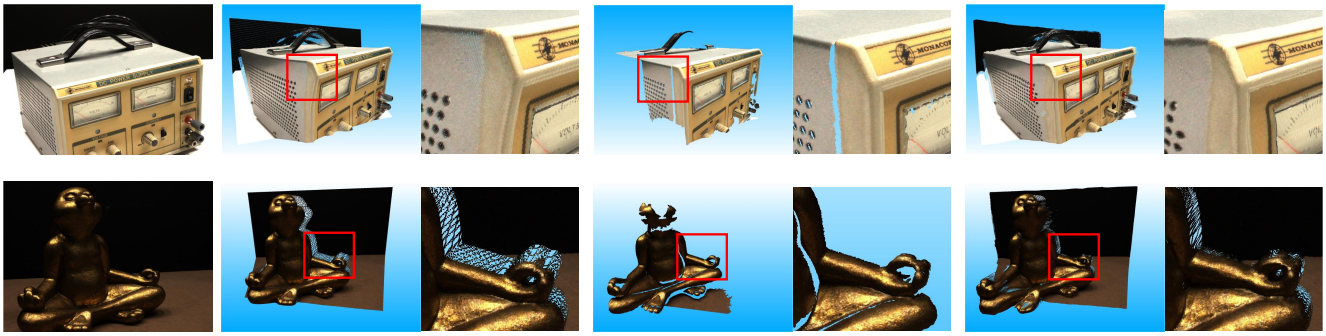

Figure 1. Comparison of monocular depth maps constructed by DA and PDA methods. The monocular depth maps and their corresponding ground-truth depth map are reprojected into 3D reconstructions to facilitate detailed comparisons. Compared to the DA method, the PDA method exhibits better scale consistency, albeit with slightly inferior surface smoothness.

1. Analysis of Monocular Depth Map

We employ COLMAP[4] to generate sparse depth maps for each view and further integrate the Prior Depth Anything (PDA)[5] method to construct corresponding monocular depth maps. In contrast, prior approaches that combine monocular depth estimation typically rely on Depth Anything V2 (DA)[7] to generate monocular depth maps. However, since the depth maps estimated by DA lack scale consistency, these methods utilize the sparse depth maps to compute the corresponding scaling factors and offsets for aligning the depth maps. In this section, we compare the aforementioned two methods, analyzing their monocular depth estimation performance and corresponding reconstruction outcomes on the DTU dataset. Specifically, we also leverage the sparse depth maps obtained from COLMAP to align the monocular depth maps generated by DA. We reproject both the monocular depth maps of individual images and the ground-truth depth maps into 3D space for comparative evaluation. The comparison results are shown in Figure 1. Evidently, both types of methods produce reliable monocular depth maps that provide relatively accurate spatial position information. However, the depth maps constructed by the DA method still exhibit scale inconsistencies even after alignment, whereas the PDA-derived depth maps demonstrate strong scale consistency, closely matching the ground-truth. Additionally, the DA method shows superior surface smoothness. This explains why we adopt the DA method in the MDGE module to provide monocular features for the reference image.

To fully demonstrate the differences among various monocular depth estimation methods, we show depth fusion results from two adjacent views processed with different monocular depth estimation and alignment methods in Fig 2. We explored three different schemes: **a)** Predicting depth maps using DepthAnythingV2(DA2) and normalizing them with the depth range; **b)** Obtaining sparse depth maps for corresponding views via COLMAP, followed by aligning the monocular depth map; **c)** Constructing monocular depth maps using Prior DepthAnything(PDA) with sparse depth maps from COLMAP. As shown in Fig 2, method **a** showed poor scale alignment; Method **b** significantly improved alignment but introduced noticeable local distortions; Method **c** further enhanced alignment, yet its surface smoothness is inferior to that of method **b**. Additionally, the recently proposed DepthAnythingV3 [3] (DA3, published after the submission of this paper) supports multi-view depth map estimation with given camera poses. The accuracy and scale consistency of the predicted depth maps are significantly better than those of monocular depth estimation methods, as shown in the part **d** in Fig 2. This performance inspired us. In our future work, We will further explore how to integrate the results of these methods as monocular depth priors with the MVS framework.

Reconstruction Performance. We separately evaluate the reconstruction performance on the DTU dataset when using these two methods to generate monocular depth maps for constructing SPE, as well as the performance of using monocular depth maps alone for 3D reconstruction. The results are shown in Table 1, indicating that using monocular depth maps alone yields significantly inferior 3D re-

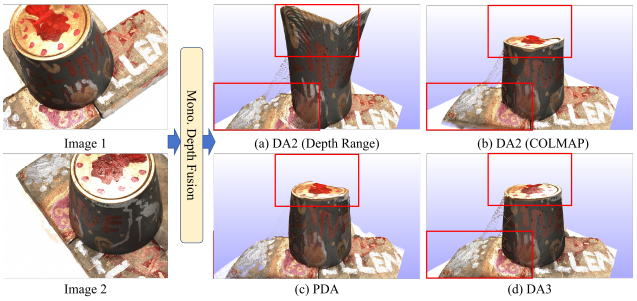

Figure 2. Visual comparison of depth fusion from two images using different monocular depth estimation and alignment methods.

construction quality compared to methods integrating MVS. Notably, when directly employing monocular depth maps for 3D reconstruction, PDA demonstrates a clear advantage over DA, with the overall metric improving from 2.586 to 1.020. This validates the superiority of scale-consistent depth maps compared to aligned depth maps. Therefore, we adopted the PDA method to construct SPE. Compared to using the DA method, the overall metric is optimized from 0.275 to 0.272.

Table 1. Ablation study for monocular depth estimation methods.

Method	Overall↓	Acc.↓	Comp.↓
DA only	2.586	4.166	1.006
PDA only	1.020	1.690	0.349
DA + SPE-MVS	0.275	0.326	0.224
PDA + SPE-MVS	**0.272**	**0.324**	**0.220**

Efficiency Comparison. We further compare our method against other state-of-the-art (SOTA) approaches on the DTU dataset in terms of efficiency, with the results summarized in Table 2. All methods were evaluated under a unified configuration using five input images at a resolution of 832 × 1152, with comparisons made on two aspects: runtime and GPU memory consumption. We also present the efficiency performance of our method at full resolution. The results show that while our method outperforms GoMVS in efficiency, it still lags behind MonoMVSNet. In future work, we will attempt to further optimize the efficiency of our method.

Table 2. Efficiency comparison with SOTA methods at a resolution of 832 × 1152 (* represents the resolution of 1152 × 1600).

Methods	Time(s)↓	Mem.(GB)↓
GoMVS [6]	0.64	12.6
MVSFormer++ [1]	**0.23**	4.7
MonoMVSNet [2]	0.25	**2.0**
Ours	0.31	3.0
Ours*	0.56	5.4

MDGE Module Details. In the MDGE module, we incorporate both 2D CNN layers and 3D CNN layers within both MFE and MDE components. Therefore, during the processing, the input of the module must undergo corresponding dimensional transformations. Specifically, in the 2D CNN-based branch, we treat the depth dimension of the probability maps $\boldsymbol{P}^k_{init}$ and $\boldsymbol{P}^k_f \in \mathbb{R}^{1\times Z_k\times H\times W}$ as the channel dimension, yielding a processed shape of $\mathbb{R}^{Z_k\times H\times W}$. Furthermore, by setting the number of convolutional kernels in the final 2D CNN layer of this branch to Z_k, the output of this branch has a dimensionality of $\mathbb{R}^{Z_k\times H\times W}$. In the 3D CNN-based branch, the last 3D CNN layer is configured with a single convolutional kernel, so that its final output has a dimensionality of $\mathbb{R}^{1\times Z_k\times H\times W}$. Subsequently, we transform the dimensions of the output from the 2D CNN-based branch into $\mathbb{R}^{1\times Z_k\times H\times W}$, thereby achieving dimensional alignment with the results from both branches.

2. More Visualization Results

We demonstrate all the reconstructed point clouds of our method on DTU and Tanks & Temples datasets, as shown in Figures 3 and 4 .

References

[1] Chenjie Cao, Xinlin Ren, and Yanwei Fu. Mvsformer++: Revealing the devil in transformer's details for multi-view stereo. In *Proceedings of the International Conference on Learning Representations*, pages 1–14, 2024. 2

[2] Jianfei Jiang, Qiankun Liu, Haochen Yu, Hongyuan Liu, Liyong Wang, Jiansheng Chen, and Huimin Ma. Monomvsnet: Monocular priors guided multi-view stereo network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 27806–27816, 2025. 2

[3] Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025. 1

[4] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4104–4113, 2016. 1

[5] Zehan Wang, Siyu Chen, Lihe Yang, Jialei Wang, Ziang Zhang, Hengshuang Zhao, and Zhou Zhao. Depth anything with any prior, 2025. 1

[6] 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 Conference on Computer Vision and Pattern Recognition*, pages 20207–20216, 2024. 2

[7] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Proceedings of the Advances in Neural Information Processing Systems*, 37:21875–21911, 2024. 1

Figure 3. Visualization of reconstructed point clouds for each scene on the DTU dataset.

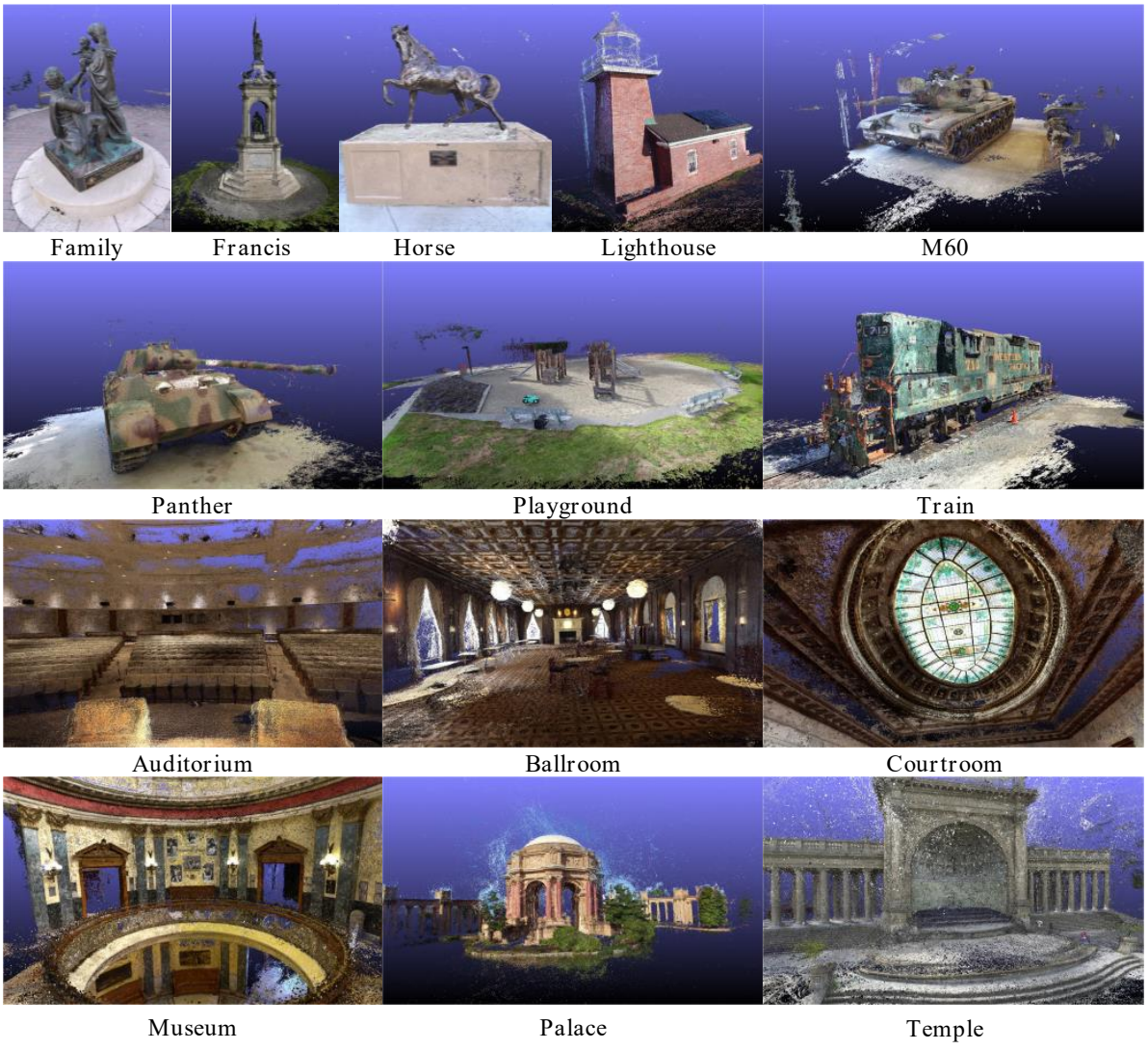

Figure 4. Visualization of reconstructed point clouds for each scene on the Tanks and Temples dataset.