FLARE: Feed-forward Geometry, Appearance and Camera Estimation from Uncalibrated Sparse Views
−−Supplementary Material−−

Appendix

In this supplementary material, we include: (1) extended implementation details covering data processing and training strategies, (2) additional experimental results, and (3) a comprehensive description of the network architecture.

A. Implementation Details

Data Processing. We follow the processing protocol of DUSt3R to generate point maps for most datasets. However, the DL3DV dataset only provides the annotation for camera parameters. To include DL3DV into our training framework, we use the multi-view stereo algorithm from COLMAP to annotate per-frame depth maps, which are then converted into point maps. Additionally, we utilize multi-view photometric and geometric consistency to eliminate noisy depth [15]. For the datasets captured as video sequences, we randomly select 8 images from a single video clip, with each video clip containing no more than 250 frames. For multi-view image datasets, we randomly select 8 images per scene.

Baselines for Novel View Synthesis. We compare our novel view synthesis results with MVSplat [3], pixelSplat [2], and CoPoNeRF [7] on the DL3DV dataset [9]. However, these methods were originally trained on only 2 views and perform not well under our sparse-view setting of 8 views. To ensure a fair comparison, we selected the two source views closest to the target rendering view as inputs for these baselines (e.g., MVSplat). We found that selecting two closest source views significantly improved their rendering quality compared to using all 8 views directly.

Numbers of Input Image. We have two camera latents: one for the first image (reference), and one is shared by all other images (source). The source token is duplicated N-1 times. Therefore, the model can process any number of input images.

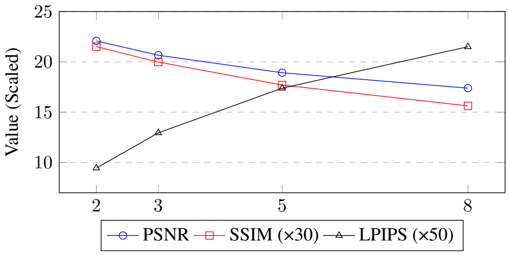

Figure 1. **Relationship between MVSplat Performance and Input Views.**

B. Experiments

Relationship between MVSplat Performance and Input Views. We evaluated MVSplat with two views because its performance degrades with additional input frames, as shown in Fig. 1, as demonstrated in the figure above. We therefore reported its optimal results.

Table 1. **Performance with a Varying Number of Input Frames.** We study the impact of changing the number of input views on the performance of our method on the DTU dataset.

Metric	2 Views	6 Views	10 Views	16 Views	25 Views
AUC@30° ↑	59.09	70.45	80.15	81.52	**81.81**
ACC. ↓	4.07	2.79	0.94	**0.24**	0.30

Study between Performance and the Number of Frames. We analyzed the impact of varying the number of frames on pose and point map estimation using the DTU dataset. For this experiment, we randomly selected 2, 6, 10, 16, and 25 source views while fixing two query views for testing pose accuracy and for evaluating surface accuracy. Under the 2-view setting, our method generates a reasonable shape, but its precision remains limited. The results demonstrate that increasing the number of views leads to improvements in both pose and surface accuracy. However, these improvements gradually plateau as the number of views continues to grow, as shown in Tab. 1.

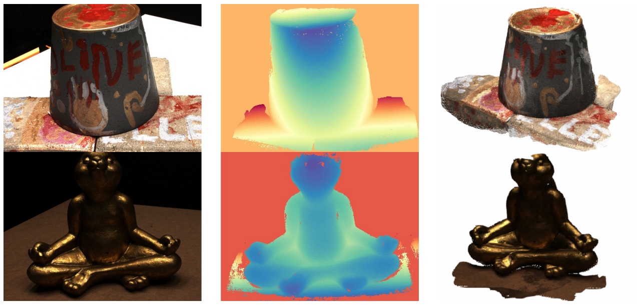

Figure 2. **Qualitative Visualization of Sparse-view 3D Reconstruction on the DTU dataset.** We visualize the input image (left), depth map (middle), and point cloud (right).

Table 2. **Dense View 3D Reconstruction on the DTU dataset.** We compare our method with baseline approaches using accuracy, completeness, and overall metrics under the dense view setting.

Methods	Accuracy↓	Completion↓	Overall↑
Camp [1]	0.835	0.554	0.695
Furu [5]	0.613	0.941	0.777
Tola [11]	0.342	1.190	0.766
Gipuma [6]	0.283	0.873	0.578
MVSNet [17]	0.396	0.527	0.462
CVP-MVSNet [16]	0.296	0.406	0.351
UCS-Net [4]	0.338	0.349	0.447
CER-MVS [10]	0.359	0.305	0.332
CIDER [14]	0.417	0.437	0.427
PatchmatchNet [12]	0.427	0.377	0.417
GeoMVSNet [18]	0.331	0.259	0.295
MASt3R [8]	0.403	0.344	0.374
DUSt3R [13]	2.677	0.805	1.741
Ours	**1.932**	**0.715**	**1.321**

Dense View 3D Reconstruction on the DTU dataset. We present the results for dense view 3D reconstruction on the DTU dataset in Tab. 2, although dense reconstruction is not our primary objective. As observed, our method achieves better results compared to DUSt3R but falls short of MASt3R. This is expected since our approach is not tailored for dense reconstruction, whereas MASt3R is specifically optimized for it through the training of matching heads.

Visualization of Sparse-view Reconstruction. We present the visualizations of our sparse-view reconstruction results on the DTU dataset in Fig. 2.

References

[1] Neill D. F. Campbell, George Vogiatzis, Carlos Hernández, and Roberto Cipolla. Using multiple hypotheses to improve depth-maps for multi-view stereo. In *ECCV*, 2008. 2

[2] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19457–19467, 2024. 1

[3] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. *arXiv preprint arXiv:2403.14627*, 2024. 1

[4] 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 *CVPR*, 2020. 2

[5] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *PAMI*, 2010. 2

[6] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *ICCV*, 2015. 2

[7] Sunghwan Hong, Jaewoo Jung, Heeseong Shin, Jiaolong Yang, Seungryong Kim, and Chong Luo. Unifying correspondence, pose and nerf for pose-free novel view synthesis from stereo pairs. *arXiv preprint arXiv:2312.07246*, 2023. 1

[8] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. *arXiv preprint arXiv:2406.09756*, 2024. 2

[9] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, Xuanmao Li, Xingpeng Sun, Rohan Ashok, Aniruddha Mukherjee, Hao Kang, Xiangrui Kong, Gang Hua, Tianyi Zhang, Bedrich Benes, and Aniket Bera. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22160–22169, 2024. 1

[10] Zeyu Ma, Zachary Teed, and Jia Deng. Multiview stereo with cascaded epipolar raft. In *ECCV*, 2022. 2

[11] Engin Tola, Christoph Strecha, and Pascal Fua. Efficient large-scale multi-view stereo for ultra high-resolution image sets. *Mach. Vis. Appl.*, 2012. 2

[12] Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Pablo Speciale, and Marc Pollefeys. Patchmatchnet: Learned multi-view patchmatch stereo. In *CVPR*, pages 14194–14203, 2021. 2

[13] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 2

[14] Qingshan Xu and Wenbing Tao. Learning inverse depth regression for multi-view stereo with correlation cost volume. In *AAAI*, 2020. 2

[15] 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. 1

[16] Jiayu Yang, Wei Mao, José M. Álvarez, and Miaomiao Liu. Cost volume pyramid based depth inference for multi-view stereo. In *CVPR*, pages 4876–4885, 2020. 2

[17] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *ECCV*, 2018. 2

[18] Zhe Zhang, Rui Peng, Yuxi Hu, and Ronggang Wang. Geomvsnet: Learning multi-view stereo with geometry perception. In *CVPR*, 2023. 2