SelfSplat: Pose-Free and 3D Prior-Free Generalizable 3D Gaussian Splatting

Supplementary Material

A. Additional Details

A.1. Architectural Details

For the prediction of 3D Gaussians [9], we utilize the monocular, multi-view encoder and the fusion block. Unlike previous methods that utilize DepthAnything [26] as a monocular encoder [25, 28] or UniMatch [24] as a multi-view encoder [4, 19], we only employ the encoder part of Croco [23] as our monocular encoder which is trained in a fully self-supervised manner. For the multi-view encoder, we adopt the backbone of [24] with randomly initialized weights. Then, we unify features from monocular and multi-view encoders using DPT [15] block. For a detailed architecture for the fusion module, see Fig. 1.

A.2. Implementation Details

For our monocular encoder, we utilized Adapter [2], which keeps the model parameters frozen while training additional residual networks for each layer. Specifically, a residual MLP block, comprising a down-projection layer and an up-projection layer, is introduced within each layer of the transformer encoder. Considering the channel dimension of the original encoder, $C^{\text{mono}} = 1024$, we set the low rank hidden dimension of AdaptMLP, $C^{\text{adapt}} = 32$, to efficiently reduce computational overhead while maintaining sufficient capacity for adaptation.

For 3D Gaussian primitives, we set the order of spherical harmonics expansion to 1, enabling the representation to extend beyond the Lambertian color model. When warping the color model from each frame's local coordinate system into an integrated global space which requires the Wigner matrices in general case, we simplify the rotation of the first level of spherical harmonics, $Y_1(r_d) = [Y_1^{-1}(r_d), Y_1^0(r_d), Y_1^1(r_d)]$, as follows:

$$Y_1(r_d) = \sqrt{\frac{3}{4\pi}} \Pi \, r_d, \quad \Pi = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix},$$

where $r_d \in \mathbb{S}^2$ is the viewing direction derived from the estimated camera poses. We adopt this warping protocol from Splatter Image [21] which is a pose-required generalizable 3D reconstruction model using 3D Gaussian Splatting.

A.3. Training Details

We train all baseline models, including ours, using custom data loaders. For RealEstate10K [29] (RE10k) and ACID [14] datasets, the distance between context frames is progressively increased from 5 to 25, and target frames are randomly selected between the context frames within this range. Each model is trained for 200K iterations and for baselines we used the default hyperparameter settings provided by the respective authors. The only exception is DBARF [3], which is trained for 400K iterations due to its official implementation supporting only a batch size of one. We provide our detailed training hyperparameters in Tab. 1 and we train our model on a singe H100 GPU, which takes approximately for 3 days. For the experiment on DL3DV [13] dataset, we initialize the model with pretrained weights from RE10k dataset and train it for 50K iterations on a single H100 GPU with a batch size of 6. The distance between context frames is gradually increased from 2 to 10. This procedure is applied to FlowCAM [18] in the same way which is the baseline model on DL3DV dataset.

For VAE [11], which was initially designed for novel view synthesis from a single image, we modify its architecture following the approach in [27] to handle multi-view input images. Specifically, we employ two separate encoders and use their mean output as the input to the decoder which synthesize novel view images. All other hyperparameters remain the same as the official implementation.

SelfSplat	
Config	Value
optimizer	Adam [10]
scheduler	Linear
learning rate	1e-4
gradient clipping	0.5
batch size	12
total iters.	200,000
warmup iters.	2,000

Table 1. Training hyperparameters.

A.4. Evaluation Details

During the evaluation on RE10k and ACID datasets, we set the interval between context frames to 40 and select the middle frame as the target view point. This target frame is used as the ground truth for metric evaluations in novel view synthesis and camera pose estimation. For the overlap categories, we utilize the pretrained feature matching model, RoMa [6], to estimate the overlap ratios between the first context frame and the target frame.

For RE10k dataset, the split proportions are 18.26% for large, 60.56% for medium, and 21.17% for small categories. In ACID dataset [14], the proportions are 33.05% for large, 41.15% for medium, and 25.80% for small.

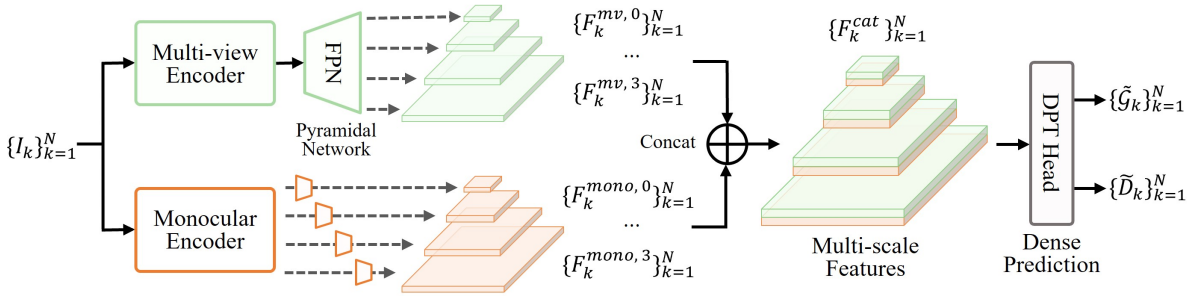

Figure 1. Detailed 3D Gaussian prediction architecture. This module takes only context images as input.

B. Additional Experiment Analysis

B.1. Inference Cost

We report the memory and time consumption required to synthesize a single 256 × 256 image during the inference stage in Table 2. Memory usage is measured as the peak memory during inference, while the number of rays per batch is adjusted if necessary. Except for VAE [11], which generates novel view images without rendering operations (utilize 2D CNN blocks) and thus fail to reconstruct interpretable 3D scene representations, our method achieves significantly lower memory usage and faster rendering speed with explicit 3D representations, demonstrating its efficiency and practical usage in real-world scenarios.

Method	Mem. (GB)	Time (s)
VAE [11]	4.694	**0.0003**
DBARF [3]	2.079	0.254
FlowCAM [18]	16.644	0.801
CoPoNeRF [7]	16.802	5.624
Ours	**1.795**	0.002

Table 2. Memory and time consumption analysis. All baselines including ours are measured on a single NVIDIA RTX 4090 GPU.

B.2. Using N Context Views

We further evaluate the model's performance across various numbers of input views, considering its practical application where more than two views are commonly used. The total number of frames is evenly divided based on the number of context views, and target frames are sampled between the context frames. Additionally, we generate a camera trajectory using the selected view points (context and target), and the Absolute Trajectory Error (ATE) is measured to validate the accuracy of the reconstructed camera path. We evaluate on RE10k dataset with 3 context views (80 frames) and 4 context views (120 frames) settings. As shown in Tab. 3 and Fig. 2, our method demonstrates superior performance in both novel view synthesis and camera trajectory estimation, as well as its ability to scale effectively with multiple input views and estimations over extended ranges without any further finetuning.

	3 views (80 frames)				4 views (120 frames)			
Method	PSNR↑	SSIM↑	LPIPS↓	ATE↓	PSNR↑	SSIM↑	LPIPS↓	ATE↓
FlowCAM [18]	19.75	0.630	0.412	0.048	18.91	0.606	0.449	0.081
Ours	**21.12**	**0.761**	**0.241**	**0.031**	**19.52**	**0.717**	**0.283**	**0.053**

Table 3. Quantitative results of using different numbers of context views on RE10k dataset.

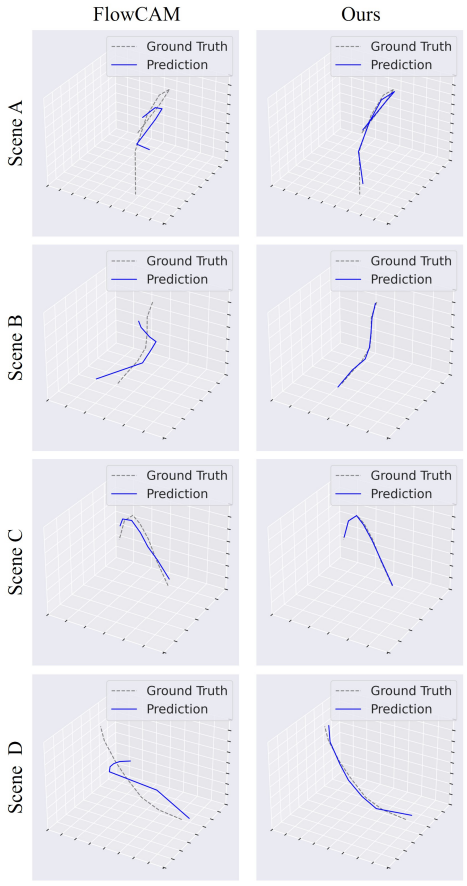

Figure 2. Visualization of camera trajectory on RE10k dataset. Construction of trajectory only consider the translation part of the estimated camera poses.

B.3. Additional Comparison

For the reader's reference, we provide a comparison with Splatt3R [17], which is also a pose-free, feed-forward Gaussian Splatting method for 3D reconstruction and novel view synthesis from stereo pairs. We omitted this model in the main paper because it requires ground-truth depth and camera pose annotations during training, which are not available in the datasets we used: RE10k, ACID [14], and DL3DV [13]. Acknowledging the differences in training data—Splatt3R was trained on ScanNet++ [5], whereas our model was trained on RE10k—we evaluate them on the DTU [8] dataset, which is an out-of-distribution dataset for both. As shown in Tab. 4 and Fig. 3, our method achieves better performance than the baseline in both evaluation metrics and visual quality, and also outperforms pixelSplat [1] which is a pose-required method in training and evaluation stage. The main reason Splatt3R cannot estimate a consistent scene scale is its reliance on a fixed pretrained MASt3R [12] model, which is trained using metric camera poses, and difference between estimated intrinsic parameters and ground truth intrinsic parameters. Thus, using the DTU dataset, which consists of unseen novel scenes, they fail to align consistent 3D Gaussians.

Training Data	Method	PSNR↑	SSIM↑	LPIPS↓
ScanNet++ [5]	Splatt3R [17]	10.24	0.295	0.629
RE10k [29]	pixelSplat [1]	12.89	0.382	0.560
	MVSplat [4]	13.94	0.473	0.385
	Ours	**13.14**	**0.425**	**0.448**

Table 4. Quantitative results of novel view synthesis on DTU dataset. While pixelSplat [1] and MVSplat [4] are pose-required methods, we include them for the reader's reference.

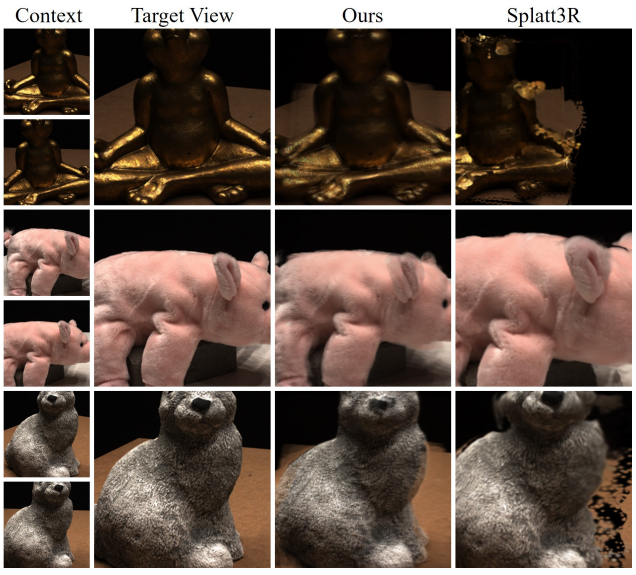

Figure 3. Qualitative comparison of novel view synthesis on DTU dataset.

B.4. Baseline Comparisons

We provide additional baseline results on cross-dataset generalization in Tab. 5.

	RE10k → ACID				ACID → RE10k			
Method	PSNR ↑	SSIM ↑	Rot.° ↓	Trans.° ↓	PSNR ↑	SSIM ↑	Rot.° ↓	Trans.° ↓
VAE	23.67	0.649	1.619	117.721	18.98	0.537	3.744	65.884
DBARF	23.56	0.644	1.772	51.969	12.60	0.502	3.306	47.851
Ours	**26.60**	**0.793**	**1.119**	**18.607**	**21.65**	**0.728**	**1.618**	**17.993**

Table 5. Additional comparison on cross-dataset generalization.

B.5. Additional Ablation and Analysis

We provide additional ablation studies and analyses, focusing on our encoder module. All methods are trained on RE10k [29] for 50k iterations, following the same procedure as in the main paper. As shown in Tab. 6, our feature fusion module with CroCo [23] initialization shows superior results in evaluation metrics.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	Rot. Avg.° ↓	Trans. Avg.° ↓
Ours	**22.65**	**0.764**	**0.222**	**1.036**	**13.705**
No CroCo [23] Init	22.15	0.738	0.240	1.091	14.042
No Monocular Encoder	21.88	0.733	0.247	1.394	17.125
No Multi-view Encoder	21.47	0.731	0.243	1.233	16.327

Table 6. Ablation studies on the encoder module design.

Pretrained weight. Since our goal is to use only unposed raw video datasets without 3D priors, we utilized CroCo, trained in a self-supervised manner. While DUSt3R [22] or MASt3R [12] pre-trained weight could enhance performance, we focus on demonstrating that 3D foundation models can be trained without costly 3D annotations.

B.6. Architectural and Evaluation Design

We designed our evaluation protocol assuming that there are no given poses, so we made a separate pose block (context and target) and a Gaussian branch (only context) independently. Thus, target images are used to estimate camera poses for following novel view synthesis evaluations. All baselines follow this protocol in their original implementations, except for CoPoNeRF [7] which utilizes given camera poses, so we substitute these poses in CoPoNeRF with estimated ones for a fair comparison.

B.7. Depth Visualization

We also provide the visualization of depth maps generated through rendering, which is essential for producing interpretable 3D representations. By comparing the results of our method with previous approaches, as shown in Fig. 4, SelfSplat demonstrates robust and reliable depth maps derived from 3D scene structures.

C. Limitations

While we demonstrate high-quality 3D geometry estimation in this work, the current framework still possesses limitations. First, further technical improvements are needed to

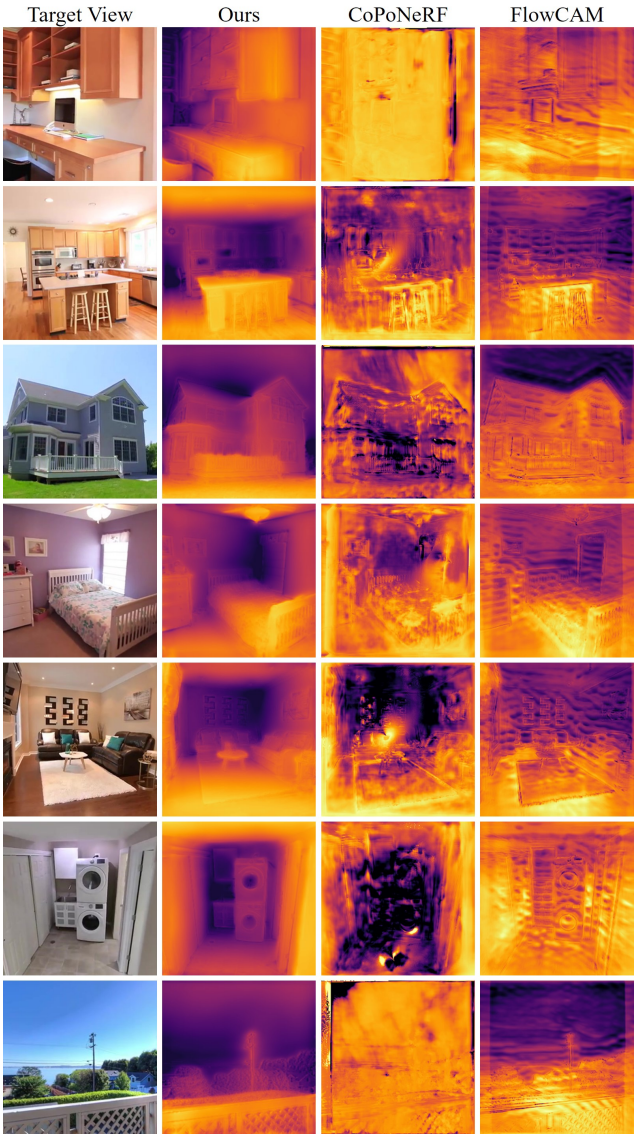

Figure 4. Qualitative comparison of depth visualization on RE10k dataset. Depth maps are obtained following the rendering process.

support wider baseline scenarios, such as a 360° scene reconstruction from unposed images in a single forward pass. Second, our framework struggles with dynamic scenes where both camera and object motion are present. Addressing these complex scenarios may benefit from incorporating multi-modal priors [16, 20] for robust and consistent alignment across wide and dynamic scene space.

D. Additional Results

We provide additional results on the following pages including novel view synthesis and epipolar line visualizations.

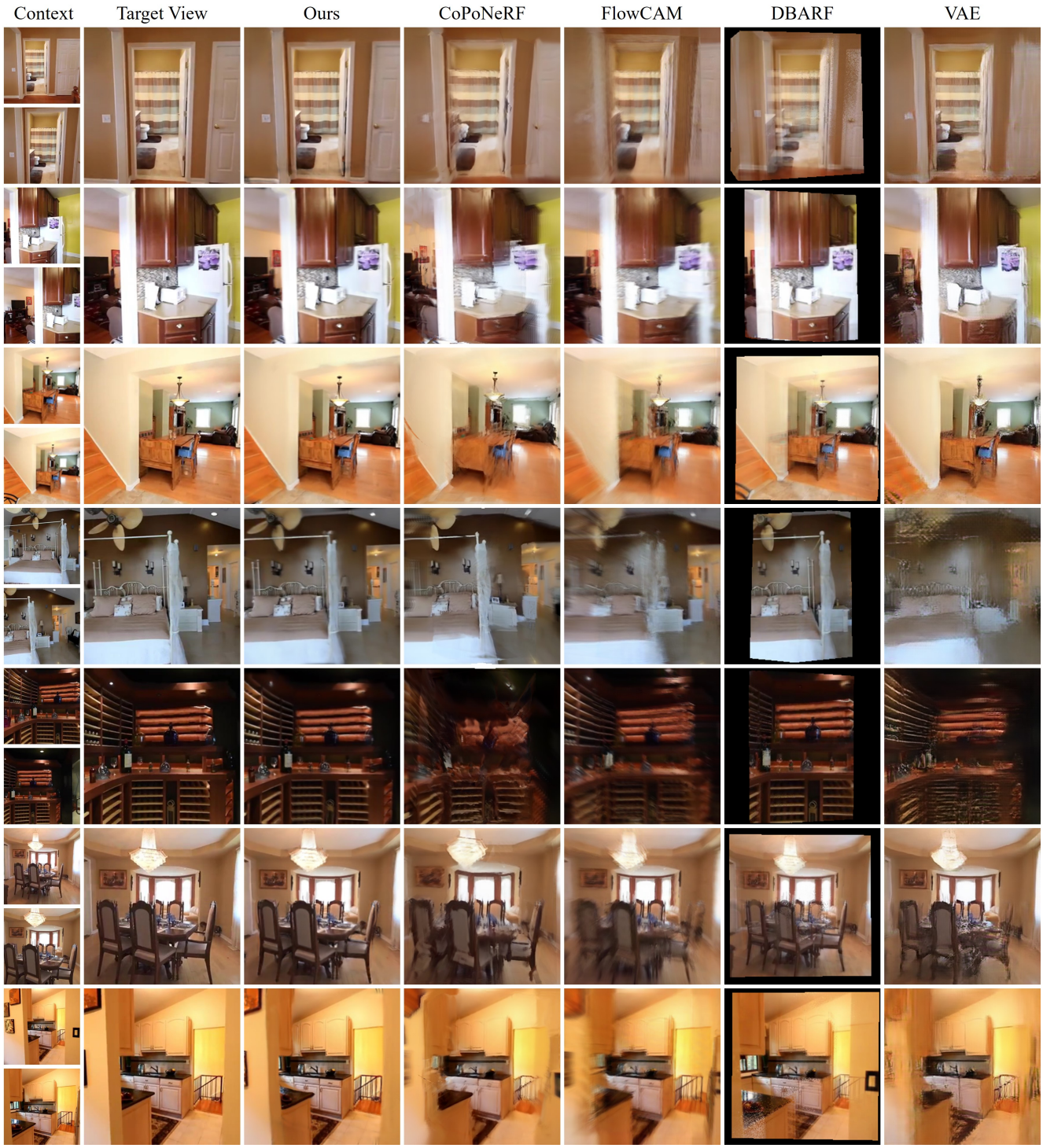

Figure 5. Qualitative comparison of novel view synthesis on RE10k dataset.

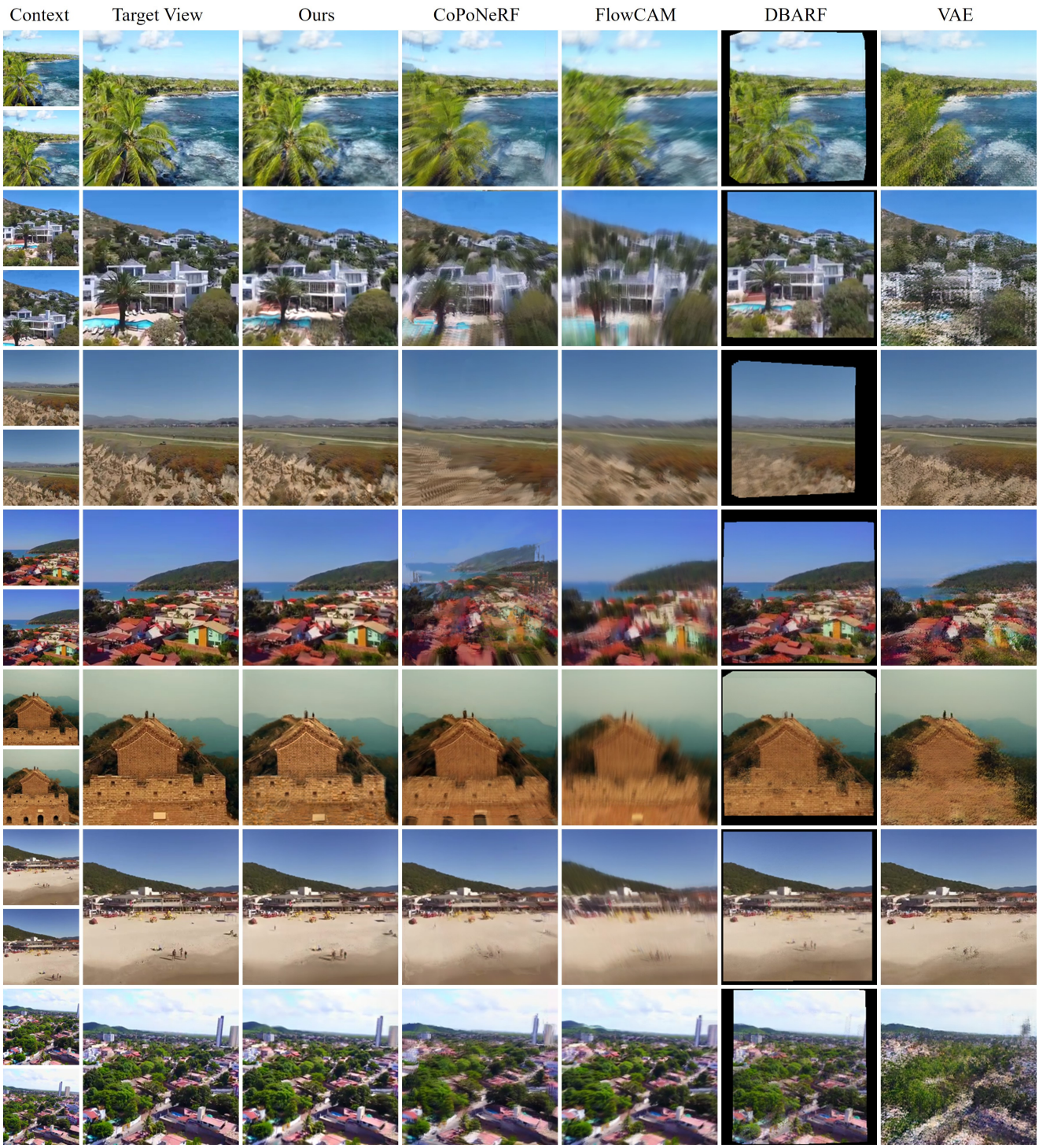

Figure 6. Qualitative comparison of novel view synthesis on ACID dataset.

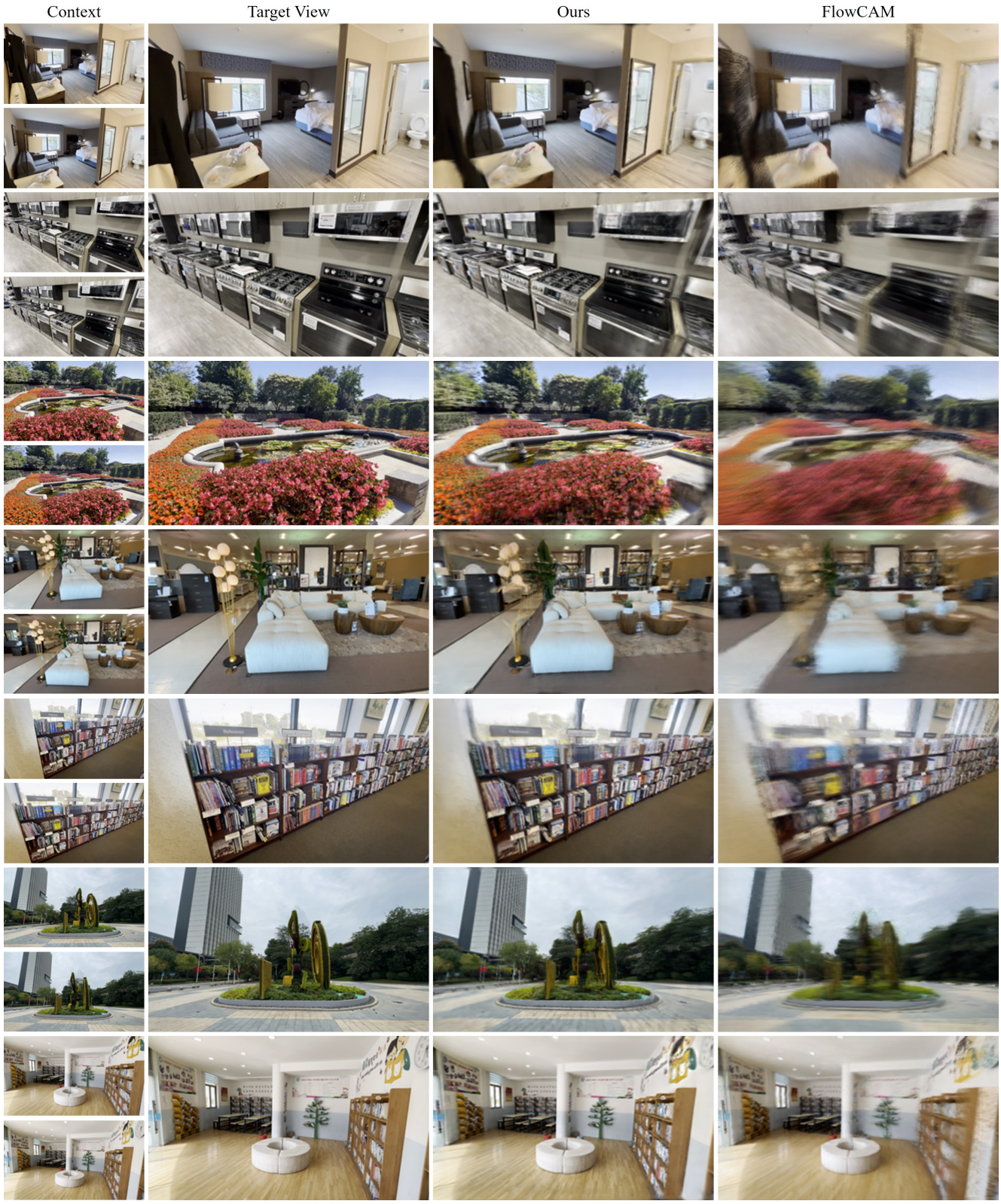

Figure 7. Qualitative comparison of novel view synthesis on DL3DV dataset.

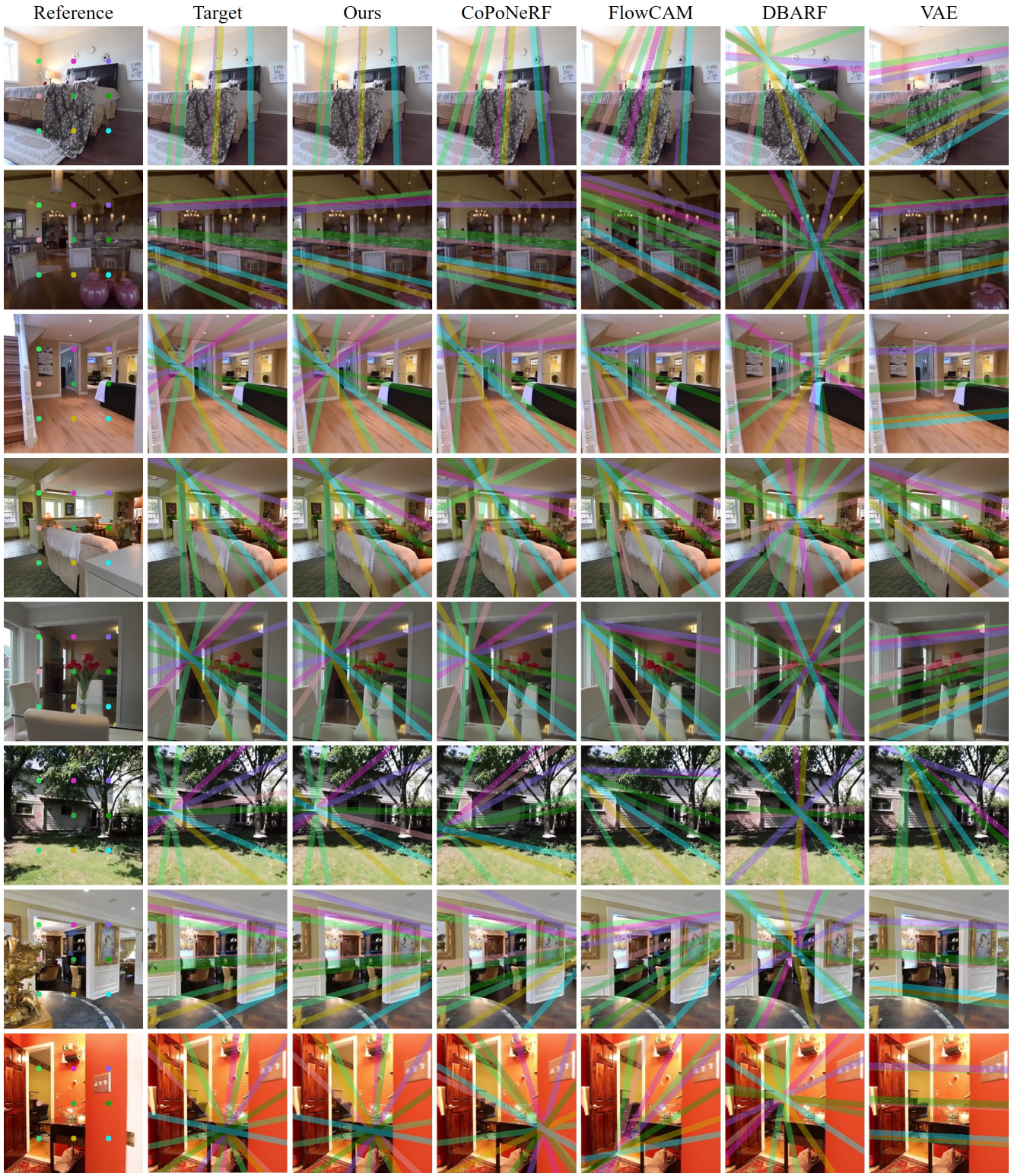

Figure 8. Epipolar lines visualization on RE10k dataset. We draw the lines from reference to target frame using relative camera pose.

References

[1] 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. 3

[2] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022. 1

[3] Yu Chen and Gim Hee Lee. Dbarf: Deep bundle-adjusting generalizable neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24–34, 2023. 1, 2

[4] 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, 3

[5] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 3

[6] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Roma: Robust dense feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19790–19800, 2024. 1

[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. 2, 3

[8] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 406–413, 2014. 3

[9] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1

[10] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[11] Zihang Lai, Sifei Liu, Alexei A Efros, and Xiaolong Wang. Video autoencoder: self-supervised disentanglement of static 3d structure and motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9730–9740, 2021. 1, 2

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

[13] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. 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*, pages 22160–22169, 2024. 1, 3

[14] Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa. Infinite nature: Perpetual view generation of natural scenes from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14458–14467, 2021. 1, 3

[15] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 1

[16] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 4

[17] Brandon Smart, Chuanxia Zheng, Iro Laina, and Victor Adrian Prisacariu. Splatt3r: Zero-shot gaussian splatting from uncalibarated image pairs. *arXiv preprint arXiv:2408.13912*, 2024. 3

[18] Cameron Smith, Yilun Du, Ayush Tewari, and Vincent Sitzmann. Flowcam: training generalizable 3d radiance fields without camera poses via pixel-aligned scene flow. *arXiv preprint arXiv:2306.00180*, 2023. 1, 2

[19] Libo Sun, Jia-Wang Bian, Huangying Zhan, Wei Yin, Ian Reid, and Chunhua Shen. Sc-depthv3: Robust self-supervised monocular depth estimation for dynamic scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2023. 1

[20] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023. 4

[21] Stanislaw Szymanowicz, Chrisitian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10208–10217, 2024. 1

[22] Shuzhe Wang et al. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 3

[23] Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Jérôme Revaud. Croco: Self-supervised pre-training for 3d vision tasks by cross-view completion. *Advances in Neural Information Processing Systems*, 35:3502–3516, 2022. 1, 3

[24] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1

[25] Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. Depthsplat: Connecting gaussian splatting and depth. *arXiv preprint arXiv:2410.13862*, 2024. 1

[26] 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

[27] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4578–4587, 2021. 1

[28] Chuanrui Zhang, Yingshuang Zou, Zhuoling Li, Minmin Yi, and Haoqian Wang. Transplat: Generalizable 3d gaussian splatting from sparse multi-view images with transformers. *arXiv preprint arXiv:2408.13770*, 2024. 1

[29] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018. 1, 3