E-RayZer: Self-supervised 3D Reconstruction as Spatial Visual Pre-training

Supplementary Material

Overview

This supplementary material is organized as follows:

- Section A: Additional implementation details.
- Section B: Details on supervised finetuning.
- Section C: Additional details on curriculum learning ablations.
- Section D: Analysis of E-RayZer trained with pose supervision.
- Section E: Additional results where E-RayZer is used as pre-training for the VGGT* model, with comparisons to RayZer [25].
- Section F: Further analysis of the training data.
- Section G: Extended qualitative comparisons with baseline methods.

A. Additional Implementation Details

This section includes more implementation details.

Training. E-RayZer is trained on 8 A100 GPUs with a global batch size of 192 (24 per GPU) for 152K iterations, taking approximately 198 hours. During the first 86K iterations, the learning curriculum progresses linearly along different sequence-sampling metrics – geometric (default) and semantic visual overlap, as well as frame interval – as described in Sec. 4.4. The learning rate schedule includes a 3K-iteration linear warm-up to a peak of 4e-4, followed by cosine decay to zero til the end of training. We use the AdamW optimizer (β_1=0.9, β_2=0.95) with gradient clipping at 1.0, and skip optimization steps when the gradient norm exceeds 5.0 before clipping.

For our 7-dataset model (Sec. 4.1), we train on a mixture of datasets with the following sampling ratios: DL3DV [35]: 1.0, CO3Dv2 [44]: 0.25, RealEstate10K [85]: 0.5, MVImgNet [77]: 0.25, ARKitScenes [6]: 0.5, WildRGB-D [68]: 0.25, and ACID [36]: 0.5. These ratios follow a simple heuristic: we downweight object-centric datasets and assign a slightly larger weight to DL3DV, which offers the most diverse and high-quality samples.

Experiments on supervised finetuning are conducted on 8 A100 GPUs as well, but with a smaller global batch size of 96. The finetuning stage runs for 50K iterations.

Architecture. E-RayZer uses a patch size of 16 and an image resolution of 256. As described in Sec. 3.2, we replace RayZer's [25] vanilla global attention with VGGT's [59] local-global alternating transformer layers for both pose estimation (f_{θ}^{cam}) and scene reconstruction ($f_{\psi'}^{\text{scene}}$). Both modules use 8 layers, each composed of one global attention layer and one frame-attention layer. Our feature dimension is 768, and we use 12 attention heads. For image and Plücker ray map tokenization, as well as for the Gaussian decoder ($f_{\omega}^{\text{gauss}}$), we simply use a single linear layer.

For a fair comparison with RayZer, all RayZer models used in this paper are trained with our proposed curriculum and the improved architecture.

Evaluation. For pose estimation and novel-view synthesis, we use fixed sequence lengths for the test sequences of each dataset and sample views with equal temporal spacing. Following RayZer, we ensure that the first and last images of each sequence are always included in the reference set. The sequence lengths are as follows: WildRGB-D [68]: 96 (Tab. 1) and 192 (Tab. 2), ScanNet++ [75]: 48, DL3DV [35]: 96, RealEstate10K [85]: 256, CO3Dv2 [44]: 96, 7-Scenes [50]: 256, Cambridge Landmarks [30]: 96, BlendedMVS [71]: 24, and NAVI [22]: 24. For (training and) evaluating pairwise flow prediction on StaticThings3D [49], we adopt the pre-computed image pairs provided by the DUSt3R [64] GitHub repository.

B. More Details on Supervised Finetuning

Here we provide additional details on the supervised finetuning experiments in Sec. 4.3.

Supervised Finetuning with E-RayZer. E-RayZer's backbone does not distinguish between the first view and the other views in the input, as it adopts a pairwise pose estimation strategy (see Sec. 3.2). In contrast, supervised pose estimation typically assumes a first-view coordinate frame (*e.g.*, DUSt3R [64] and VGGT [59]). To incorporate this inductive bias into our backbone, we introduce an additional camera token dedicated to the first image (in addition to the existing learned camera token) and train it from scratch. The camera tokens are processed by E-RayZer's pose estimation module (f_{θ}^{cam}) and subsequently passed to VGGT's camera head for supervised pose estimation. For depth estimation and pairwise flow prediction, the DPT head takes as input the intermediate feature maps generated by the Gaussian-based scene reconstruction module ($f_{\psi'}^{\text{scene}}$). For E-RayZer and all other baselines, the DPT head uses four feature maps extracted from equally spaced transformer layers. Note that our Gaussian-based scene reconstruction module takes the predicted reference-view Plücker ray maps as input, but only in the pose and depth estimation experiments are the predicted camera poses supervised. For pairwise flow prediction, the predicted poses produced by the pose head remain unsupervised to ensure a fair comparison with other baselines.

Details on Other Baselines. For baselines that use different spatial or temporal patch sizes (*e.g.*, E-RayZer uses a tem-

Table 7. **Comparison with a Pose-supervised Baseline on Novel-view Synthesis (NVS) and Pose Estimation.** We report PSNR for NVS and RPA$_\uparrow$@5°/15°/30° for pose estimation. While the pose-supervised baseline generally outperforms the self-supervised model on coarse pose accuracy (RPA$_\uparrow$@15°/30°), its novel-view synthesis quality is consistently lower.

Method	Training Data	NAVI [22]				ScanNet++ [75]				DL3DV [35]			
		PSNR$_\uparrow$	@5°$_\uparrow$	@15°$_\uparrow$	@30°$_\uparrow$	PSNR$_\uparrow$	@5°$_\uparrow$	@15°$_\uparrow$	@30°$_\uparrow$	PSNR$_\uparrow$	@5°$_\uparrow$	@15°$_\uparrow$	@30°$_\uparrow$
Pose-sup. Baseline	DL3DV [35]	13.4	12.8	51.1	**72.5**	16.7	4.4	**33.7**	**64.5**	15.0	**78.1**	**94.7**	**97.8**
E-RayZer (ours)		**20.5**	**20.7**	**57.8**	69.6	**20.1**	**7.7**	33.6	63.0	**20.3**	72.0	88.4	93.5
Pose-sup. Baseline	7 datasets	13.5	18.9	**61.6**	**80.6**	17.3	**6.4**	**35.7**	**67.4**	14.9	53.0	**85.0**	**93.2**
E-RayZer (ours)		**20.6**	**24.6**	56.1	69.2	**20.7**	5.7	34.8	63.7	**19.7**	**59.9**	82.9	90.2

Table 8. **Comparison with RayZer [25] as a Pre-trained Backbone.** The top block reports results for models trained on DL3DV [35], and the bottom block reports results for models trained on a mixture of seven datasets. Note that pre-training and supervised finetuning are performed on the same data (*i.e.*, DL3DV or the 7-dataset mixture). We report pose accuracy RPA$_\uparrow$@5°/15°. Models are labeled as self-supervised or supervised. VGGT* denotes our re-implementation with E-RayZer's pairwise camera head. The top-three results are color-ranked from red to yellow. E-RayZer provides stronger pre-training than RayZer.

	Method	DL3DV [35]		RE10K [85]		CO3Dv2 [44]		WildRGB-D [68]		7-Scenes [50]		CamLand [29]		BlendedMVS [71]		NAVI [22]		ScanNet++ [75]	
		@5°$_\uparrow$	@15°$_\uparrow$	@5°$_\uparrow$	@15°$_\uparrow$	@5°$_\uparrow$	@15°$_\uparrow$	@5°$_\uparrow$	@15°$_\uparrow$	@5°$_\uparrow$	@15°$_\uparrow$	@5°$_\uparrow$	@15°$_\uparrow$	@5°$_\uparrow$	@15°$_\uparrow$	@5°$_\uparrow$	@15°$_\uparrow$	@5°$_\uparrow$	@15°$_\uparrow$
DL3DV	RayZer [25]	0.0	0.6	0.0	0.2	0.0	0.6	0.0	0.0	0.0	0.2	0.0	0.3	0.0	0.5	0.0	0.6	0.0	0.7
	E-RayZer (ours)	72.0	88.4	83.0	96.8	19.1	61.8	51.1	82.3	38.8	78.0	18.1	62.9	22.9	46.8	20.7	57.8	7.7	33.6
	VGGT*	79.6	94.2	80.4	97.9	16.0	64.3	32.5	76.2	34.7	83.6	11.1	49.8	17.0	42.8	14.3	54.5	6.7	39.8
	RayZer→VGGT*	84.4	95.3	85.7	98.4	24.9	71.2	43.9	86.4	38.0	83.6	27.3	73.0	24.0	45.8	25.5	58.3	12.2	49.6
	E-RayZer→VGGT*	87.3	96.6	85.3	98.4	25.3	72.2	56.2	91.4	43.8	82.8	30.2	75.6	29.2	52.2	26.9	64.3	14.3	53.8
7 datasets	RayZer [25]	0.0	1.9	0.0	0.9	0.0	1.6	0.0	1.1	0.0	2.0	0.0	0.6	0.0	1.6	0.0	1.6	0.0	0.9
	E-RayZer (ours)	59.9	82.9	84.1	97.5	30.3	74.2	63.1	85.3	26.0	76.5	9.8	47.3	22.3	45.5	24.6	56.1	5.7	34.8
	VGGT*	66.1	88.9	85.2	98.5	43.4	83.5	76.8	96.0	31.1	78.0	22.9	66.3	19.0	49.9	28.8	67.3	13.1	54.8
	RayZer→VGGT*	72.8	91.7	88.1	98.6	53.8	85.1	81.5	96.3	37.7	84.9	28.3	65.7	24.3	52.7	34.6	70.4	15.0	58.7
	E-RayZer→VGGT*	78.8	92.8	91.0	99.1	58.9	86.3	86.4	96.7	42.7	88.3	35.2	64.4	31.5	57.7	41.5	73.7	22.0	65.2

poral batch size of 1, whereas VideoMAE V2 [61] uses 2), we first resize or repeat the input so that the number of output tokens matches that of our model. For these methods, we generally adopt the "base" model checkpoints provided in their official GitHub repositories, as they roughly match the computational budget of our model.

C. Additional Details on Curriculum Ablation

In this section, we provide additional details on the baseline setups used in Tab. 6. We compare our visual-overlap-based curricula to two baseline strategies: (1) Non-curriculum baseline, where we do not progressively increase the difficulty of training samples. Concretely, the geometric visual-overlap score remains fixed within the range [0.5, 1.0] throughout training, without any linear decay. As a result, the model encounters challenging samples (*e.g.*, wide-baseline views) from the very beginning. (2) Frame-interval-based curriculum, where geometric-overlap scores are converted into frame intervals that linearly increase over training. To construct the interval schedule for each dataset, we pre-sample 10K sequences with geometric-overlap scores in [0.5, 1.0] and set the maximum frame interval to the 95th percentile of these sequences. This heuristic implicitly defines dataset-specific hyperparameters that would otherwise need to be *manually tuned*.

D. A Pose-supervised Baseline

We introduce a pose-supervised baseline whose pose estimation module is trained using ground-truth camera poses (typically obtained from running Structure-from-Motion systems [48]), following prior supervised methods (*e.g.*, DUSt3R [64] and VGGT [59]). In this baseline, the Gaussian-based scene reconstruction module is still optimized with a photometric loss; however, gradients from this loss are not propagated back to the pose estimation module. The results are shown in Tab. 7.

We observe that while the pose-supervised baseline usually outperforms E-RayZer on coarse pose accuracy (RPA@15°/30°), it consistently achieves lower PSNR for novel-view synthesis. We attribute this weaker NVS performance to a misalignment between the predicted poses and the Gaussian prediction. To supervise pose estimation, the ground-truth camera poses are normalized to a predefined scale (*e.g.*, 1.0), and the pose estimation module learns to predict camera poses at this scale. However, the Gaussian prediction module does not necessarily follow the same scale. In practice, we observe many training instances where the rendered Gaussians fall outside the image plane, providing little or no useful photometric supervision.

In contrast, with our curriculum design, E-RayZer learns pose estimation and Gaussian prediction jointly, allowing both components to automatically align to the same scale. This avoids the scale-misalignment issue and leads to more

Table 9. **Additional Results on Data Mixing and Scaling.** We train E-RayZer with different combinations of datasets. Compared to Tab. 5, we additionally include SpatialVID [60], a large in-the-wild video dataset. Results are color-ranked from red to yellow. Mixing datasets improves distribution coverage, whereas simply using larger datasets does not necessarily yield better performance – both diversity and data quality play critical roles.

Training Data	# Seq.	NAVI [22]				CO3Dv2 [44]				ScanNet++ [75]				DL3DV [35]			
		PSNR↑	@5°↑	@15°↑	@30°↑	PSNR↑	@5°↑	@15°↑	@30°↑	PSNR↑	@5°↑	@15°↑	@30°↑	PSNR↑	@5°↑	@15°↑	@30°↑
RE10K [85]	66K	17.2	1.8	16.9	34.0	19.1	0.6	8.3	26.0	17.5	1.1	13.3	37.3	17.3	21.2	55.0	72.7
SpatialVID [60]	100K	17.9	0.7	11.2	26.4	19.9	0.2	5.7	20.9	18.0	0.3	6.7	26.0	17.2	11.4	36.6	56.0
DL3DV [35]	10K	20.5	20.7	57.8	69.6	22.9	19.1	61.8	78.8	20.1	7.7	33.6	63.0	20.3	72.0	88.4	93.5
7-dataset Mix	352K	20.6	24.6	56.1	69.2	24.3	30.3	74.2	83.7	20.7	5.7	34.8	63.7	19.7	59.9	82.9	90.2

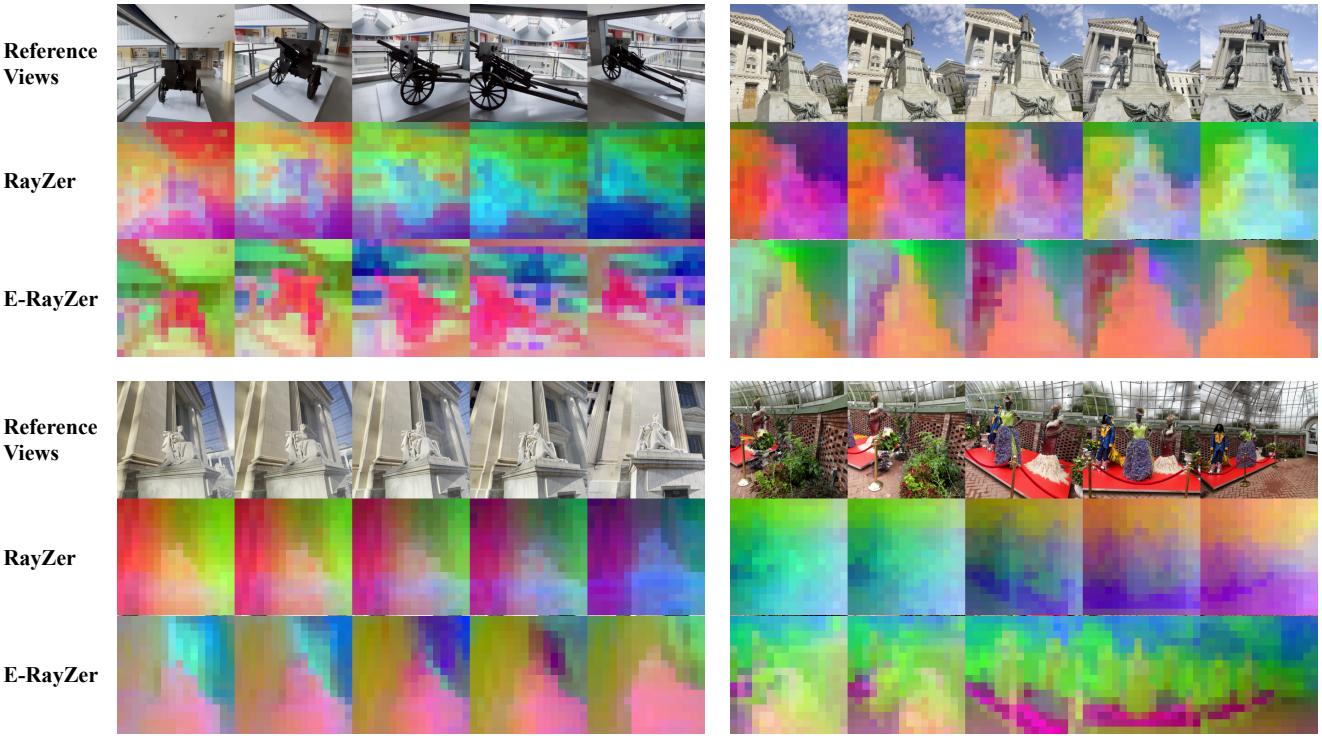

Figure 5. **Additional Visual Comparison with RayZer [25] on Learned Features.** We visualize feature maps using their top-three PCA components. The features produced by E-RayZer exhibit stronger and more spatially consistent patterns that align well with the underlying scene structure, whereas RayZer's features show noticeable color shifts across frames.

stable training and stronger novel-view synthesis performance. In short, this experiment further confirms the benefit of our self-supervised 3D reconstruction framework for both camera pose estimation and novel-view synthesis.

E. Additional Results on Pre-training

We present additional results where E-RayZer is used as a pre-trained backbone for VGGT* (our re-implementation of VGGT [59], matched to our architecture and training data). We compare E-RayZer against RayZer [25] as an alternative pre-training approach and evaluate pose accuracy across multiple datasets.

Tab. 8 summarizes results under two training configurations: using only DL3DV [35] and using a mixture of seven datasets. Note that pre-training and supervised fine-tuning are conducted on the same data (*i.e.*, DL3DV or the 7-dataset mixture). In both settings, VGGT* initialized with E-RayZer outperforms its RayZer-initialized counterpart on most metrics, indicating that the representations learned by E-RayZer provide stronger and more transferable pre-training for downstream supervised pose estimation.

F. Further Analysis of Training Data

We further analyze how different training datasets affect model performance.

Compared to Tab. 5, Tab. 9 additionally includes E-RayZer results on a static subset of SpatialVID [60], a large in-the-wild video dataset, and reports the number of training sequences used in each setting. We observe that a larger number of training sequences does not necessarily yield higher performance. For example, the model trained on 100K SpatialVID sequences performs comparably to the

RealEstate10K [85] model (which uses 66K sequences), yet significantly underperforms the DL3DV [35] model (which contains only 10K sequences). We conjecture that this gap stems from the noisy nature of in-the-wild data: SpatialVID sequences originate primarily from internet videos, and our training subsets are selected using their coarse dynamic-ratio labels. Also, SpatialVID often features simple or near-static camera motions. In contrast, DL3DV is carefully curated without moving objects and contains high-quality video sequences with diverse camera trajectories. These results support our earlier observations about data quality and highlight the importance of data curation when scaling self-supervised learning to large in-the-wild resources.

We also find that mixing datasets improves distribution coverage and leads to better generalization. For instance, models trained with mixed data perform better on the object-centric CO3Dv2 [44] compared to models trained solely on non-object-centric datasets.

Finally, we note that all experiments are conducted under a fixed computation budget (*i.e.*, 152K iterations with a global batch size of 192). Within this controlled setting, our results consistently suggest that diversity and quality of data matter more than quantity for training self-supervised models. We believe that collecting diverse, high-quality data remains both a key challenge and a promising direction for future work.

G. More Qualitative Comparisons

Learned Feature Representations. In Fig. 5, we provide additional qualitative results comparing the learned feature representations of E-RayZer with those of RayZer [25]. The feature maps produced by E-RayZer exhibit more stable and coherent patterns across views, while RayZer's feature maps often display noticeable color shifts between frames. These results suggest that E-RayZer learns feature representations that are more geometrically grounded.

Pose Estimation and Novel-view Synthesis. We present additional qualitative comparison with baselines in Fig. 6. Compared to SPFSplat [21], E-RayZer consistently achieves better pose accuracy and higher-quality novel-view synthesis, despite being trained entirely from scratch without relying on pretrained priors such as MASt3R [33]. RayZer [25] generally produces high-quality novel views; however, it often exhibits grid-like artifacts in uncertain regions (highlighted with red bounding boxes). Moreover, RayZer's predicted poses are not physically aligned with the scene, whereas the camera poses learned by E-RayZer are geometrically grounded.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022. 3

[2] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025. 2

[3] Vassileios Balntas, Shuda Li, and Victor Prisacariu. Relocnet: Continuous metric learning relocalisation using neural nets. In *ECCV*, 2018. 2

[4] Mohamed El Banani, Jason J Corso, and David F Fouhey. Novel object viewpoint estimation through reconstruction alignment. In *CVPR*, 2020. 2

[5] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024. 3

[6] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. In *NeurIPS D&B*, 2021. 5, 1

[7] Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, et al. Perception encoder: The best visual embeddings are not at the output of the network. In *NeurIPS*, 2025. 1, 2, 7

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020. 2

[9] Ruojin Cai, Bharath Hariharan, Noah Snavely, and Hadar Averbuch-Elor. Extreme rotation estimation using dense correlation volumes. In *CVPR*, 2021. 2

[10] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 3

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 3

[12] Yu Chen and Gim Hee Lee. Dbarf: Deep bundle-adjusting generalizable neural radiance fields. In *CVPR*, 2023. 3

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 2

[14] Bardienus Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jerome Revaud. Mast3r-sfm: a fully-integrated solution for unconstrained structure-from-motion. In *3DV*, 2025. 2, 6

[15] Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. Masked autoencoders as spatiotemporal learners. In *NeurIPS*, 2022. 3

[16] Negar Foroutan, Paul Teiletche, Ayush Kumar Tarun, and Antoine Bosselut. Revisiting multilingual data mix-

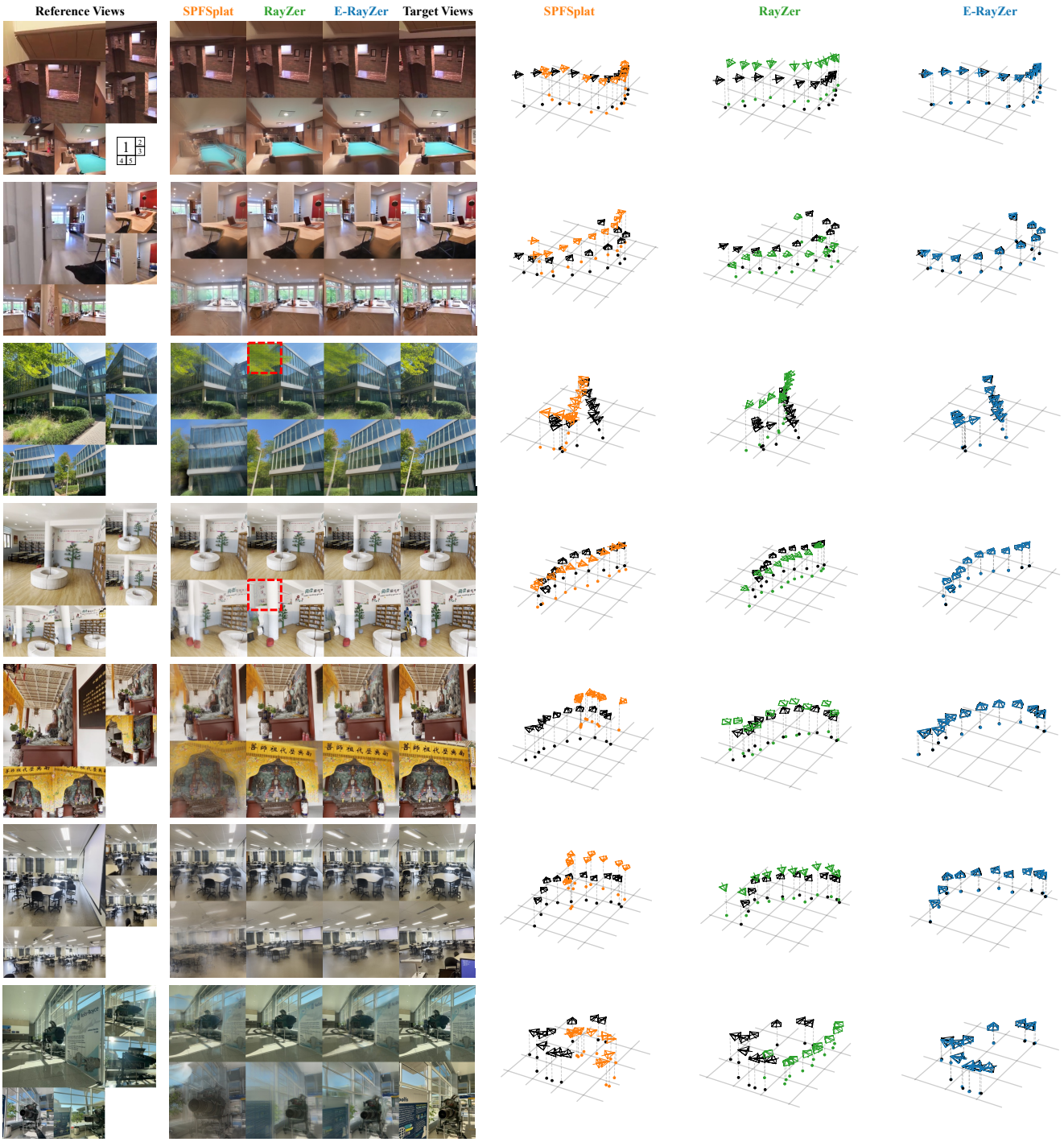

Figure 6. **Additional Visual Comparison with (Partially) Self-supervised Methods.** We show results for both novel-view synthesis (left) and pose estimation (right). The temporal order of the reference views is shown in the first row. Ground-truth poses are visualized in black, and predicted poses are aligned to the ground truth via an optimal similarity transform. E-RayZer outperforms baselines in pose accuracy, demonstrating its grounded 3D understanding. While RayZer [25] typically produces high-quality novel views, it often exhibits grid-like artifacts in low-texture regions (highlighted with red boxes; best viewed when zoomed in), likely due to its latent-rendering formulation.

tures in language model pretraining. *arXiv preprint arXiv:2510.25947*, 2025. 8

[17] Yang Fu, Ishan Misra, and Xiaolong Wang. Mononerf: Learning generalizable nerfs from monocular videos without camera poses. In *ICML*, 2023. 3

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross

Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 3

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 2, 3

[20] Sunghwan Hong, Jaewoo Jung, Heeseong Shin, Jisang Han, Jiaolong Yang, Chong Luo, and Seungryong Kim. Pf3plat: Pose-free feed-forward 3d gaussian splatting for novel view synthesis. In *ICML*, 2025. 2, 6

[21] Ranran Huang and Krystian Mikolajczyk. No pose at all: Self-supervised pose-free 3d gaussian splatting from sparse views. In *ICCV*, 2025. 2, 6, 4

[22] Varun Jampani, Kevis-Kokitsi Maninis, Andreas Engelhardt, Arjun Karpur, Karen Truong, Kyle Sargent, Stefan Popov, André Araujo, Ricardo Martin Brualla, Kaushal Patel, et al. Navi: Category-agnostic image collections with high-quality 3d shape and pose annotations. In *NeurIPS*, 2023. 7, 8, 1, 2, 3

[23] Hanwen Jiang, Zhenyu Jiang, Kristen Grauman, and Yuke Zhu. Few-view object reconstruction with unknown categories and camera poses. In *3DV*, 2024. 2

[24] Hanwen Jiang, Zhenyu Jiang, Yue Zhao, and Qixing Huang. LEAP: Liberate sparse-view 3d modeling from camera poses. In *ICLR*, 2024. 2

[25] Hanwen Jiang, Hao Tan, Peng Wang, Haian Jin, Yue Zhao, Sai Bi, Kai Zhang, Fujun Luan, Kalyan Sunkavalli, Qixing Huang, et al. Rayzer: A self-supervised large view synthesis model. In *ICCV*, 2025. 1, 2, 3, 4, 6, 7, 8, 5

[26] Lihan Jiang, Yucheng Mao, Linning Xu, Tao Lu, Kerui Ren, Yichen Jin, Xudong Xu, Mulin Yu, Jiangmiao Pang, Feng Zhao, et al. Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. In *ACM SIGGRAPH Asia*, 2025. 2

[27] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 3

[28] Gyeongjin Kang, Jisang Yoo, Jihyeon Park, Seungtae Nam, Hyeonsoo Im, Sangheon Shin, Sangpil Kim, and Eunbyung Park. Selfsplat: Pose-free and 3d prior-free generalizable 3d gaussian splatting. In *CVPR*, 2025. 2

[29] Alex Kendall and Roberto Cipolla. Geometric loss functions for camera pose regression with deep learning. In *CVPR*, 2017. 7, 2

[30] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *ICCV*, 2015. 1

[31] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. In *ACM ToG*, 2023. 2, 3, 4

[32] Zihang Lai, Sifei Liu, Alexei A Efros, and Xiaolong Wang. Video autoencoder: self-supervised disentanglement of static 3d structure and motion. In *ICCV*, 2021. 3

[33] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *ECCV*, 2024. 6, 4

[34] Amy Lin, Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose++: Recovering 6d poses from sparse-view observations. In *3DV*, 2024. 2

[35] 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 *CVPR*, 2024. 5, 6, 7, 8, 1, 2, 3, 4

[36] 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 *ICCV*, 2021. 5, 1

[37] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 3

[38] Thomas W Mitchel, Hyunwoo Ryu, and Vincent Sitzmann. True self-supervised novel view synthesis is transferable. In *ICLR*, 2026. 3

[39] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. In *TMLR*, 2024. 2, 3, 5, 7

[40] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016. 3

[41] Julius Plucker. Xvii. on a new geometry of space. In *Philosophical Transactions of the Royal Society of London*, 1865. 3

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3

[43] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021. 7

[44] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *ICCV*, 2021. 5, 7, 8, 1, 2, 3, 4

[45] Chris Rockwell, Justin Johnson, and David F Fouhey. The 8-point algorithm as an inductive bias for relative pose prediction by vits. In *3DV*, 2022. 2

[46] Mehdi SM Sajjadi, Aravindh Mahendran, Thomas Kipf, Etienne Pot, Daniel Duckworth, Mario Lučić, and Klaus Greff. Rust: Latent neural scene representations from unposed imagery. In *CVPR*, 2023. 3

[47] Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, et al. Zeronvs: Zero-shot 360-degree view synthesis from a single real image. In *CVPR*, 2024. 5

[48] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 2

[49] Philipp Schröppel, Jan Bechtold, Artemij Amiranashvili, and Thomas Brox. A benchmark and a baseline for robust multi-view depth estimation. In *3DV*, 2022. 6, 7, 1

[50] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *CVPR*, 2013. 7, 1, 2

[51] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 1, 2, 3, 7

[52] Samarth Sinha, Jason Y Zhang, Andrea Tagliasacchi, Igor Gilitschenski, and David B Lindell. Sparsepose: Sparse-view camera pose regression and refinement. In *CVPR*, 2023. 2

[53] Brandon Smart, Chuanxia Zheng, Iro Laina, and Victor Adrian Prisacariu. Splatt3r: Zero-shot gaussian splatting from uncalibrated image pairs, 2024. arXiv preprint arXiv:2408.13912. 2

[54] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *NeurIPS*, 2022. 2, 3

[55] Michael Tschannen, Manoj Kumar, Andreas Steiner, Xiaohua Zhai, Neil Houlsby, and Lucas Beyer. Image captioners are scalable vision learners too. In *NeurIPS*, 2023. 3

[56] Khiem Vuong, Anurag Ghosh, Deva Ramanan, Srinivasa Narasimhan, and Shubham Tulsiani. Aerialmegadepth: Learning aerial-ground reconstruction and view synthesis. In *CVPR*, 2025. 2

[57] Haoru Wang, Kai Ye, Yangyan Li, Wenzheng Chen, and Baoquan Chen. The less you depend, the more you learn: Synthesizing novel views from sparse, unposed images without any 3d knowledge. In *ICLR*, 2026. 3

[58] Jianyuan Wang, Christian Rupprecht, and David Novotny. Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. In *ICCV*, 2023. 2

[59] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, 2025. 1, 2, 4, 6, 7, 8, 3

[60] Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Yi Zhang, Chang Zeng, Yanxi Zhou, et al. Spatialvid: A large-scale video dataset with spatial annotations, 2025. arXiv preprint arXiv:2509.09676. 3

[61] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *CVPR*, 2023. 1, 2, 7, 8

[62] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *CVPR*, 2025. 2

[63] Ruoyu Wang, Yi Ma, and Shenghua Gao. Recollection from pensieve: Novel view synthesis via learning from uncalibrated videos. *arXiv preprint arXiv:2505.13440*, 2025. 3

[64] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 2, 1

[65] 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. In *NeurIPS*, 2022. 3

[66] Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. Croco v2: Improved cross-view completion pre-training for stereo matching and optical flow. In *ICCV*, 2023. 1, 2, 3, 7, 8

[67] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *CVPR*, 2020. 3

[68] Hongchi Xia, Yang Fu, Sifei Liu, and Xiaolong Wang. Rgbd objects in the wild: Scaling real-world 3d object learning from rgb-d videos. In *CVPR*, 2024. 5, 6, 7, 1, 2

[69] Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy S Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. Doremi: Optimizing data mixtures speeds up language model pretraining. In *NeurIPS*, 2023. 8

[70] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024. 5

[71] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *CVPR*, 2020. 6, 7, 1, 2

[72] Botao Ye, Sifei Liu, Haofei Xu, Xueting Li, Marc Pollefeys, Ming-Hsuan Yang, and Songyou Peng. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. In *ICLR*, 2025. 2

[73] Jiasheng Ye, Peiju Liu, Tianxiang Sun, Jun Zhan, Yunhua Zhou, and Xipeng Qiu. Data mixing laws: Optimizing data mixtures by predicting language modeling performance. In *ICLR*, 2025. 8

[74] Vickie Ye, Ruilong Li, Justin Kerr, Matias Turkulainen, Brent Yi, Zhuoyang Pan, Otto Seiskari, Jianbo Ye, Jeffrey Hu, Matthew Tancik, et al. gsplat: An open-source library for gaussian splatting. In *JMLR*, 2025. 4

[75] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *ICCV*, 2023. 6, 7, 8, 1, 2, 3

[76] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *CVPR*, 2021. 2

[77] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, et al. Mvimgnet: A large-scale dataset of multi-view images. In *CVPR*, 2023. 5, 1

[78] Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose: Predicting probabilistic relative rotation for single objects in the wild. In *ECCV*, 2022. 2

[79] Jason Y. Zhang, Amy Lin, Moneish Kumar, Tzu-Hsuan Yang, Deva Ramanan, and Shubham Tulsiani. Cameras as rays: Sparse-view pose estimation via ray diffusion. In *ICLR*, 2024. 2, 3

[80] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting. In *ECCV*, 2024. 2

[81] Yuchen Zhang, Nikhil Keetha, Chenwei Lyu, Bhuvan Jhamb, Yutian Chen, Yuheng Qiu, Jay Karhade, Shreyas Jha, Yaoyu Hu, Deva Ramanan, et al. Ufm: A simple path towards unified dense correspondence with flow. In *NeurIPS*, 2025. 5, 7

[82] Qitao Zhao and Shubham Tulsiani. Sparse-view pose estimation and reconstruction via analysis by generative synthesis. In *NeurIPS*, 2024. 2

[83] Qitao Zhao, Amy Lin, Jeff Tan, Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Diffusionsfm: Predicting structure and motion via ray origin and endpoint diffusion. In *CVPR*, 2025. 2

[84] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017. 3

[85] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. In *ACM SIGGRAPH*, 2018. 5, 6, 7, 8, 1, 2, 3, 4

[86] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *CVPR*, 2023. 2