VGG-T^3: Offline Feed-Forward 3D Reconstruction at Scale

Supplementary Material

In this appendix, we provide:

- **A detailed description of the implementation and training process** of VGG-T^3, including dataset usage, image collection sampling (based on co-visibility), training hyperparameters, and the specific parameters that are optimized during training (Sec. A).
- **Enhancements to the VGGT baseline** that enables processing of larger image collections as well as *increased accuracy* making it a stronger baseline. (Sec. B)
- **Further ablation studies** on key components of our method, including the effect of the number of optimizer steps used for the Test-Time Training (TTT) objective and an investigation into different filter configurations for the ShortConv2D layer, showing optimal settings (Sec. C).
- **Additional qualitative results and visualizations** for comparison with baselines (VGGT, TTT3R), qualitative examples of visual localization, and a discussion of the method's performance on scenes with larger spatial extent (Sec. D).

A. Implementation Details

Training. We list the datasets used for training in Sec. A. To obtain an image collection during training, we follow a greedy sampling approach: The algorithm starts by randomly sampling the first image, then uniformly samples from the set of images with co-visibility greater than 0.3 with any of the images currently in the collection. This step repeats until the desired collection size is reached. We precompute the required co-visibility matrix via a depth consistency check [85].

Following VGGT, we use an adaptive batch size with image collections of 2-24 images while keeping the total number of images per GPU at approximately 48. The image aspect ratio is sampled uniformly from the interval $[0.5, 2.0]$, and images are then resized such that their longer side is 518. During training, we apply color jitter augmentation to each image independently, making the network more robust to brightness and contrast changes.

We train VGG-T^3 using AdamW [61] with a learning rate of 10^{-4}, weight decay of 0.05, and $\beta_1 = 0.9, \beta_2 = 0.95$. The learning rate increases by a factor of 10 during the first $1,000$ training steps, then decays following a cosine schedule to a final learning rate of 10^{-6}. For the inner optimization of the test-time training objective, we use Muon [45] with 5 Newton-Schulz iterations, a learning rate of 0.1, and employ 1 optimizer step during training. The TTT MLPs use input and output dimension 1024, matching the hidden state size of VGGT, and projects to $4\times$ the input dimension in their hidden layers. We train only the QKV projection matrices as well as the output projection in the global attention layers and the newly introduced parameters of the TTT module, while keeping all remaining parameters of the VGGT architecture (including encoder, per-image attention, and prediction heads) frozen.

Additionally, only the values projected from image patch tokens participate in the ShortConv2D operation. The camera and register tokens are passed through.

Type	Dataset
Indoor	Aria Synthetic Environments [5]
	DynamicReplica [47]
	Hypersim [73]
	Replica [83]
	Cubify Anything [54]
	Scannet++ [109]
	Scannet [18]
	Taskonomy [110]
Outdoor	Mapillary Metropolis [72]
	MatrixCity [56]
	Megadepth [57]
	Mid-Air [29]
	Mapillary Planet-scale Depth Dataset [2]
	ParallelDomain4D [25, 90]
	vKITTI2 [14, 31]
Object centric	CO3Dv2 [71]
	Kubric [33]
	Wild-RGBD [102]
Mixed	BlendedMVG [108]
	DL3DV-10K [58]
	Spring [63]
	TartanAirV2 [98]
	UnrealStereo4k [89]

Table 7. Datasets used for training.

Inference details. The VGGT architecture, which we initialize with, has multiple decoders that predict redundant geometric quantities. To obtain pointmaps one can either use the outputs of the global pointmap prediction head directly or use the camera and depth predictions together to unproject to pointmaps. While VGGT finds the latter to be more precise, we use the global prediction head to obtain pointmaps due to the imprecise camera pose predictions mentioned in Sec. 4.1, which would otherwise degrade the pointmaps obtained by unprojecting depth. In Sec. 4.3, we retain the camera tokens of all mapping images as input to the camera head in the visual localization setting since

VGGT's camera head requires the camera tokens of all images. The camera token of the query image then participates in the softmax attention operation in the camera head before it is decoded to camera parameters. For all benchmarking, we use NVIDIA A100-80GB GPUs.

Further evaluation details. For visual localization results in Sec. 4.3, we sub-sample mapping images at a stride of 200 for 7Scenes and 20 for Wayspots. For pointmap evaluation Sec. 4.2, the usage of the iterative closest point (ICP) algorithm for alignment of prediction and ground truth point clouds makes evaluation very slow when evaluating predictions on large image sets. We instead select a set of equally spaced keyframes, that capture the scene geometry, to compute pointmap metrics while we treat all other frames as supporting views. For our evalution we use 10 keyframes. For TTT3R, we provide the images in sequential order with the keyframes last such that the model has seen all images of the scene before making predictions.

B. VGGT adjustments

To enable a fair comparison with VGGT in the setting with a large number of images in Sec. 4.2, we perform several changes in the VGGT codebase that enhance its performance.

Memory-optimizations and distributed inference. First, we follow Shen et al. [79] and discard unused activations in VGGT's alternating attention module, which allows processing up to $1k$ images on a single 80GB GPU. Next, we enable context parallel inference using Ulysses [41], implemented in TransformerEngine [1], in the global attention layers. We note that the underlying attention implementation still uses FlashAttention2 [21]. While this allows VGGT to run for $2k$ images, as we show in Sec. 4.2, this requires runtimes up to 47 minutes on 2 GPUs.

Enhanced long-sequence generalization. For fair comparison on large image collections, we further adjust the scale parameter of the softmax in the global attention layers similar to the approach of Jin et al. [44], ensuring the entropy of the attention matrix stays constant. Let

$$a_{i,j} = \frac{\exp(\lambda k_i^T q_j)}{\sum_k \exp(\lambda k_k^T q_k)} \tag{6}$$

be the attention scores as used in softmax attention where $\lambda = 1/\sqrt{d}$ [91]. We instead set

$$\lambda' = \lambda \max(1.0, \log_{N_T} N), \tag{7}$$

where N_T and N are the maximum number of tokens seen during training and of the current sequence, respectively. This ensures that the scaling is the same for sequence lengths seen during training, while for larger sequence lengths the attention matrix is sharpened. Since VGGT trains using a maximum of 24 images with 518×518 resolution and uses patch size 14, we set $N_T = 24 * (518/14)^2 = 32,856$. We show improved performance using this entropy-scaling in Tab. 8 for large image collections making the VGGT baseline significantly stronger.

[1] https://docs.nvidia.com/deeplearning/transformer-engine/user-guide/api/pytorch.html#transformer_engine.pytorch.DotProductAttention

#images	250		500		750		1000	
	CD ↓	NC ↑	CD ↓	NC ↑	CD ↓	NC ↑	CD ↓	NC ↑
VGGT	0.018	**0.894**	0.025	0.876	0.040	0.864	0.041	0.855
VGGT + Entropy-scaling	**0.016**	**0.894**	**0.017**	**0.889**	**0.030**	**0.871**	**0.029**	**0.872**

Table 8. Attention entropy-scaling makes VGGT a stronger baseline on large image collections.

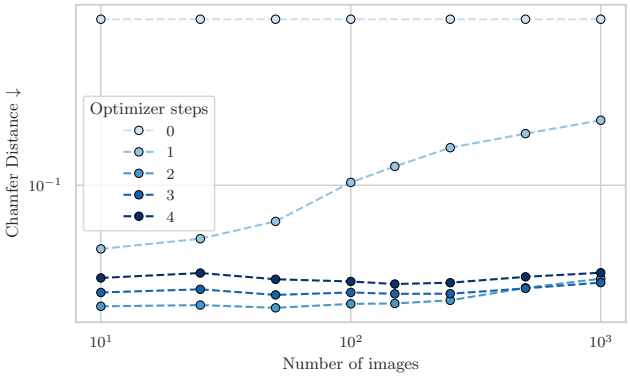


Figure 4. Pointmap error with increasing number of images when varying the optimizer steps on the TTT objective.

C. Additional Results

Number of optimizer steps. We provide an additional evaluation varying the number of steps used to optimize the TTT objective Eq. (3) at inference time for varying image collection sizes. We report results on the NRGBD dataset [6] in Fig. 4. As expected, we find that without TTT optimization the reconstruction error is high as no global information is propagated across tokens. A single optimizer step is sufficient for image collection sizes seen during training; however, the reconstruction error degrades as the number of images extends beyond that. Two optimizer steps achieve the best performance across a wide range of image collection sizes, and further increasing the number of steps to 3 or 4 leads to comparable or slightly worse performance.

ShortConv2D. In the main paper, we find improved performance when using a 3×3 ShortConv2D on values v_i of the attention operation before optimizing the MLP using TTT. Here, we provide further experiments using different configurations of our ShortConv2D. In addition to a 3×3 filter on the values v_i (V-3) used in the main paper, we consider a 5×5 filter (V-5) and a variant where we apply ShortConv2D

	CD ↓	NC ↑	mAA(30) ↑
No ShortConv2D	0.074	0.833	72.16
V-3	**0.066**	**0.838**	**74.14**
V-5	0.069	0.833	72.52
K-3	0.068	0.834	72.89
KV-3	0.081	0.820	69.44

Table 9. **Results for different filter configuration in ShortConv2D.**

to keys k_i and values v_i jointly (KV-3).

We report results in Tab. 9. We observe that increasing the filter size from 3 to 5 does not further increase performance, showing that a filter size of 3 is sufficient to obtain a strong self-supervised objective for TTT. Applying ShortConv2D to both the keys and values results in decreased performance. We explain this by the fact that applying the same spatial mixing does not break the dependency between keys and values, as explained in Sec. 4.2.

D. Additional Qualitative Results

Qualitative comparison. We report additional qualitative comparisons between VGGT, TTT3R, and VGG-T^3 in Fig. 5. TTT3R and VGG-T^3 process these $1k$ image collections within 1 minute; however, our method produces 3D consistent reconstructions while TTT3R degrades significantly. VGGT achieves slightly sharper details but takes more than 11 minutes due to the quadratic scaling of softmax attention.

Visual localization examples. Complementary to the visual localization results in Sec. 4.3, we show examples of localizing query images in the completed reconstruction by running the frozen MLPs in Fig. 6. In Fig. 7, we show an in-the-wild example where we localize a tourist picture taken from a phone camera, together with its geometry, within a recording of an autonomous vehicle from the KITTI dataset that is 7 years older. Despite the temporal gap and changes in the street, our method successfully localizes the query image. We observe that the tourist photo captures upper parts of buildings not visible from the car-mounted camera, demonstrating the robustness of our approach to viewpoint variations.

Scenes with larger spatial extent. We visualize reconstructions of Waymo sequences that have larger spatial extent in Fig. 8. While VGG-T^3 can often achieve similar results to VGGT (Fig. 8a), in some cases with more complex scene layouts, the reconstruction quality is degraded (Fig. 8b). We note this as a limitation that linear-time attention mechanisms cannot yet match softmax attention in all cases; however, this also provides an interesting avenue to explore for future work by, *e.g.*, adapting the amount of computation depending on scene complexity and designing more expressive linear attention mechanisms that match the accuracy of softmax attention.

References

[1] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. Building rome in a day. *ICCV*, 2011. 2

[2] Manuel López Antequera, Pau Gargallo, Markus Hofinger, Samuel Rota Bulò, Yubin Kuang, and Peter Kontschieder. Mapillary Planet-Scale Depth Dataset. In *CVPR*, pages 589–604, Cham, 2020. Springer International Publishing. 1

[3] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *CVPR*, 2016. 2

[4] Eduardo Arnold, Jamie Wynn, Sara Vicente, Guillermo Garcia-Hernando, Áron Monszpart, Victor Prisacariu, Daniyar Turmukhambetov, and Eric Brachmann. Map-Free Visual Relocalization: Metric Pose Relative to a Single Image. In *ECCV*, 2022. 7

[5] Armen Avetisyan, Christopher Xie, Henry Howard-Jenkins, Tsun-Yi Yang, Samir Aroudj, Suvam Patra, Fuyang Zhang, Duncan P. Frost, Luke Holland, Campbell Orme, Jakob J. Engel, Edward Miller, Richard A. Newcombe, and Vasileios Balntas. SceneScript: Reconstructing Scenes with an Autoregressive Structured Language Model. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXI*, pages 247–263. Springer, 2024. arXiv:2403.13064 [cs]. 1

[6] Dejan Azinović, Ricardo Martin-Brualla, Dan B. Goldman, Matthias Nießner, and Justus Thies. Neural RGB-D Surface Reconstruction. In *CVPR*, 2022. 6, 2

[7] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer Normalization. *arXiv preprint arXiv:1607.06450*, 2016. 3

[8] Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time. *arXiv preprint arXiv:2501.00663*, 2024. 2

[9] Gabriele Berton, Carlo Masone, and Barbara Caputo. Rethinking visual geo-localization for large-scale applications. In *CVPR*, 2022. 2

[10] Eric Brachmann and Carsten Rother. Visual camera relocalization from rgb and rgb-d images using dsac. *IEEE TPAMI*, 44(9), 2021. 2

[11] Eric Brachmann, Tommaso Cavallari, and Victor Adrian Prisacariu. Accelerated Coordinate Encoding: Learning to Relocalize in Minutes Using RGB and Poses. In *CVPR*, 2023. 7

[12] Eric Brachmann, Jamie Wynn, Shuai Chen, Tommaso Cavallari, Aron Monszpart, Daniyar Turmukhambetov, and Victor Adrian Prisacariu. Scene coordinate reconstruction: Posing of image collections via incremental learning of a relocalizer. In *ECCV*, 2024. 2, 3

[13] Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black. A Naturalistic Open Source Movie for Optical Flow Evaluation. In *ECCV*, 2012. 6

[14] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual KITTI 2, 2020. arXiv:2001.10773 [cs]. 1

[15] Yohann Cabon, Lucas Stoffl, Leonid Antsfeld, Gabriela Csurka, Boris Chidlovskii, Jerome Revaud, and Vincent Leroy. MUSt3R: Multi-view Network for Stereo 3D Reconstruction. *arXiv preprint arXiv:2503.01661*, 2025. 2

[16] Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Ttt3r: 3d reconstruction as test-time training. *arXiv preprint arXiv:2509.26645*, 2025. 2, 6

[17] Zhuoguang Chen, Minghui Qin, Tianyuan Yuan, Zhe Liu, and Hang Zhao. LONG3R: Long Sequence Streaming 3D Reconstruction. In *ICCV*, 2025. 2

[18] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Niessner. ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. In *CVPR*, 2017. 6, 1

[19] Karan Dalal, Daniel Koceja, Gashon Hussein, Jiarui Xu, Yue Zhao, Youjin Song, Shihao Han, Ka Chun Cheung, Jan Kautz, Carlos Guestrin, et al. One-minute video generation with test-time training. In *CVPR*, 2025. 2

[20] Tri Dao and Albert Gu. Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality. In *Int. Conf. Mach. Learn.*, 2024. 2

[21] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. *NeurIPS*, 35, 2022. 5, 2

[22] DeepSeek-AI. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948*, 2025. 5

[23] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, Michael Tschannen, Anurag Arnab, Xiao Wang, Carlos Riquelme Ruiz, Matthias Minderer, Joan Puigcerver, Utku Evci, Manoj Kumar, Sjoerd van Steenkiste, Gamaleldin Fathy Elsayed, Aravindh Mahendran, Fisher Yu, Avital Oliver, Fantine Huot, Jasmijn Bastings, Mark Collier, Alexey A. Gritsenko, Vighnesh Birodkar, Cristina Nader Vasconcelos, Yi Tay, Thomas Mensink, Alexander Kolesnikov, Filip Pavetic, Dustin Tran, Thomas Kipf, Mario Lucic, Xiaohua Zhai, Daniel Keysers, Jeremiah J. Harmsen, and Neil Houlsby. Scaling Vision Transformers to 22 Billion Parameters. In *Int. Conf. Mach. Learn.*, 2023. 3

[24] Kai Deng, Zexin Ti, Jiawei Xu, Jian Yang, and Jin Xie. VGGT-Long: Chunk it, Loop it, Align it – Pushing VGGT's Limits on Kilometer-scale Long RGB Sequences. *arXiv preprint arXiv:2507.16443*, 2025. 2

[25] Parallel Domain. Parallel domain. https://paralleldomain.com/, 2024. 1

[26] Siyan Dong, Shuzhe Wang, Shaohui Liu, Lulu Cai, Qingnan Fan, Juho Kannala, and Yanchao Yang. Reloc3r: Large-Scale Training of Relative Camera Pose Regression for Generalizable, Fast, and Accurate Visual Localization. In *CVPR*, 2025. 7

[27] Bardienus Pieter 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. 1

[28] Sven Elflein, Qunjie Zhou, and Laura Leal-Taixé. Light3R-SfM: Towards Feed-forward Structure-from-Motion. In *CVPR*, 2025. 2

[29] Michael Fonder and Marc Van Droogenbroeck. Mid-Air: A Multi-Modal Dataset for Extremely Low Altitude Drone Flights. In *CVPR*, pages 0–0, 2019. 1

[30] Daniel Y Fu, Tri Dao, Khaled K Saab, Armin W Thomas, Atri Rudra, and Christopher Ré. Hungry hungry hippos: Towards language modeling with state space models. *arXiv preprint arXiv:2212.14052*, 2022. 2

[31] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual Worlds as Proxy for Multi-Object Tracking Analysis. In *CVPR*, pages 4340–4349, 2016. 1

[32] A Geiger, P Lenz, C Stiller, and R Urtasun. Vision meets robotics: The KITTI dataset. *Int. Jour. of Rob. Res.*, 32(11), 2013. 6

[33] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J. Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti (Derek) Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. Kubric: A Scalable Dataset Generator. In *CVPR*, pages 3749–3761, 2022. 1

[34] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *COLM*, 2024. 2, 4

[35] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021. 2

[36] Dongchen Han, Yining Li, Tianyu Li, Zixuan Cao, Ziming Wang, Jun Song, Yu Cheng, Bo Zheng, and Gao Huang. ViT^3: Unlocking Test-Time Training in Vision, 2025. arXiv:2512.01643 [cs]. 4

[37] Stephen Hausler, Sourav Garg, Ming Xu, Michael Milford, and Tobias Fischer. Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. In *CVPR*, 2021. 2

[38] Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. Query-Key Normalization for Transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020. 3, 4

[39] Geoffrey E. Hinton and David C. Plaut. Using Fast Weights to Deblur Old Memories. *Proc. Ann. Meeting of the Cog. Sci. Soc.*, 9(0), 1987. 3

[40] Weizhe Hua, Zihang Dai, Hanxiao Liu, and Quoc Le. Transformer quality in linear time. In *Int. Conf. Mach. Learn.*, 2022. 2

[41] Sam Ade Jacobs, Masahiro Tanaka, Chengming Zhang, Minjia Zhang, Shuaiwen Leon Song, Samyam Rajbhandari, and Yuxiong He. DeepSpeed Ulysses: System Optimizations for Enabling Training of Extreme Long Sequence

Transformer Models. *arXiv preprint arXiv:2309.14509*, 2023. 2

[42] Wonbong Jang, Philippe Weinzaepfel, Vincent Leroy, Lourdes Agapito, and Jerome Revaud. Pow3R: Empowering Unconstrained 3D Reconstruction with Camera and Scene Priors. In *CVPR*, 2025. 2

[43] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanaes. Large Scale Multi-view Stereopsis Evaluation. In *CVPR*, 2014. 6

[44] Zhiyu Jin, Xuli Shen, Bin Li, and Xiangyang Xue. Training-free Diffusion Model Adaptation for Variable-Sized Text-to-Image Synthesis. *NeurIPS*, 36, 2023. 4, 2

[45] Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. 5, 1

[46] Praneeth Kacham, Vahab Mirrokni, and Peilin Zhong. Polysketchformer: Fast transformers via sketching polynomial kernels. *arXiv preprint arXiv:2310.01655*, 2023. 2

[47] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. DynamicStereo: Consistent Dynamic Depth From Stereo Videos. In *CVPR*, pages 13229–13239, 2023. 1

[48] Jungo Kasai, Hao Peng, Yizhe Zhang, Dani Yogatama, Gabriel Ilharco, Nikolaos Pappas, Yi Mao, Weizhu Chen, and Noah A. Smith. Finetuning Pretrained Transformers into RNNs. In *Proc. Emp. Met. in Nat. Lang. Proc.*, 2021. 2, 8

[49] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention. In *Int. Conf. Mach. Learn.*, 2020. 2

[50] Tong Ke and Stergios I Roumeliotis. An efficient algebraic solution to the perspective-three-point problem. In *CVPR*, 2017. 2

[51] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kontschieder. MapAnything: Universal Feed-Forward Metric 3D Reconstruction. *arXiv preprint arXiv:2509.13414*, 2025. 1, 2

[52] Laurent Kneip, Davide Scaramuzza, and Roland Siegwart. A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation. In *CVPR*, 2011. 2

[53] Yushi Lan, Yihang Luo, Fangzhou Hong, Shangchen Zhou, Honghua Chen, Zhaoyang Lyu, Shuai Yang, Bo Dai, Chen Change Loy, and Xingang Pan. STream3R: Scalable Sequential 3D Reconstruction with Causal Transformer. *arXiv preprint arXiv:2508.10893*, 2025. 2

[54] Justin Lazarow, David Griffiths, Gefen Kohavi, Francisco Crespo, and Afshin Dehghan. Cubify Anything: Scaling Indoor 3D Object Detection. In *CVPR*, pages 22225–22233, 2025. 1

[55] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding Image Matching in 3D with MASt3R. In *ECCV*, 2024. 2

[56] Yixuan Li, Lihan Jiang, Linning Xu, Yuanbo Xiangli, Zhenzhi Wang, Dahua Lin, and Bo Dai. MatrixCity: A Large-scale City Dataset for City-scale Neural Rendering and Beyond. In *ICCV*, pages 3205–3215, 2023. 1

[57] Zhengqi Li and Noah Snavely. MegaDepth: Learning Single-View Depth Prediction From Internet Photos. In *CVPR*, pages 2041–2050, 2018. 1

[58] 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 *CVPR*, pages 22160–22169, 2024. 1

[59] Hao Liu, Matei Zaharia, and Pieter Abbeel. Ring Attention with Blockwise Transformers for Near-Infinite Context. *arXiv preprint arXiv:2310.01889*, 2023. 6

[60] Yuzheng Liu, Siyan Dong, Shuzhe Wang, Yingda Yin, Yanchao Yang, Qingnan Fan, and Baoquan Chen. SLAM3R: Real-Time Dense Scene Reconstruction from Monocular RGB Videos. In *CVPR*, 2025. 2

[61] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *ICLR*. OpenReview.net, 2019. arXiv:1711.05101. 1

[62] Dominic Maggio, Hyungtae Lim, and Luca Carlone. VGGT-SLAM: Dense RGB SLAM Optimized on the SL(4) Manifold. *arXiv preprint arXiv:2505.12549*, 2025. 2

[63] Lukas Mehl, Jenny Schmalfuss, Azin Jahedi, Yaroslava Nalivayko, and Andrés Bruhn. Spring: A High-Resolution High-Detail Dataset and Benchmark for Scene Flow, Optical Flow and Stereo. In *CVPR*, pages 4981–4991, 2023. 1

[64] Jean Mercat, Igor Vasiljevic, Sedrick Keh, Kushal Arora, Achal Dave, Adrien Gaidon, and Thomas Kollar. Linearizing Large Language Models. *arXiv preprint arXiv:2405.06640*, 2024. 2, 4

[65] 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. 1

[66] Emanuele Palazzolo, Jens Behley, Philipp Lottes, Philippe Giguère, and Cyrill Stachniss. ReFusion: 3D Reconstruction in Dynamic Environments for RGB-D Cameras Exploiting Residuals. In *Int. Conf. Intell. Robot. Syst.*, 2019. 6

[67] Linfei Pan, Dániel Baráth, Marc Pollefeys, and Johannes L. Schönberger. Global Structure-from-Motion Revisited. In *ECCV*, 2024. 1

[68] Vojtech Panek, Zuzana Kukelova, and Torsten Sattler. Meshloc: Mesh-based visual localization. In *ECCV*, 2022. 2

[69] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *CVPR*, 2019. 1

[70] Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y. Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Re. Hyena Hierarchy: Towards Larger Convolutional Language Models. In *Int. Conf. Mach. Learn.*, 2023. 2, 4

[71] 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*, pages 10901–10911, 2021. 1

[72] Mapillary Research. Mapillary Metropolis Dataset. 1

[73] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A Photorealistic Synthetic Dataset for Holistic Indoor Scene Understanding. In *ICCV*, pages 10912–10922, 2021. 1

[74] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *CVPR*, 2019. 2

[75] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Efficient & effective prioritized matching for large-scale image-based localization. *IEEE TPAMI*, 39(9), 2016. 2

[76] Johannes L. Schonberger and Jan-Michael Frahm. Structure-From-Motion Revisited. In *CVPR*, 2016. 1

[77] Thomas Schops, Johannes L. Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A Multi-View Stereo Benchmark With High-Resolution Images and Multi-Camera Videos. In *CVPR*, 2017. 6

[78] Noam Shazeer. GLU Variants Improve Transformer. *arXiv preprint arXiv:2002.05202*, 2020. 5

[79] You Shen, Zhipeng Zhang, Yansong Qu, and Liujuan Cao. FastVGGT: Training-Free Acceleration of Visual Geometry Transformer. *arXiv preprint arXiv:2509.02560*, 2025. 1, 2, 5, 6

[80] 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. 2, 6, 7

[81] Noah Snavely, Steven M. Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3D. In *ACM Trans. Graph. (Proc. SIGGRAPH)*, 2006. 1

[82] Noah Snavely, Steven M. Seitz, and Richard Szeliski. Modeling the World from Internet Photo Collections. *IJCV*, 80 (2), 2008. 1

[83] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica Dataset: A Digital Replica of Indoor Spaces, 2019. arXiv:1906.05797 [cs]. 1

[84] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *Int. Conf. Intell. Robot. Syst.*, 2012. 6

[85] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-Free Local Feature Matching With Transformers. In *CVPR*, pages 8922–8931, 2021. 1

[86] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-Time Training with Self-Supervision for Generalization under Distribution Shifts. In *Int. Conf. Mach. Learn.*, 2020. 2, 3

[87] Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*, 2023. 2

[88] Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, Tatsunori Hashimoto, and Carlos Guestrin. Learning to (Learn at Test Time): RNNs with Expressive Hidden States. *arXiv preprint arXiv:2407.04620*, 2025. 1, 2, 3, 4

[89] Fabio Tosi, Yiyi Liao, Carolin Schmitt, and Andreas Geiger. SMD-Nets: Stereo Mixture Density Networks. In *CVPR*, pages 8942–8952, 2021. 1

[90] Basile Van Hoorick, Rundi Wu, Ege Ozguroglu, Kyle Sargent, Ruoshi Liu, Pavel Tokmakov, Achal Dave, Changxi Zheng, and Carl Vondrick. Generative Camera Dolly: Extreme Monocular Dynamic Novel View Synthesis. In *Computer Vision – ECCV 2024*, pages 313–331, Cham, 2025. Springer Nature Switzerland. 1

[91] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *NeurIPS*, 2017. 2, 3

[92] Chung-Shien Brian Wang, Christian Schmidt, Jens Piekenbrinck, and Bastian Leibe. Faster VGGT with Block-Sparse Global Attention. *arXiv preprint arXiv:2509.07120*, 2025. 1, 2, 5, 6

[93] Hengyi Wang and Lourdes Agapito. 3D Reconstruction with Spatial Memory. In *3DV*, 2025. 2, 6

[94] Junxiong Wang, Daniele Paliotta, Avner May, Alexander M. Rush, and Tri Dao. The Mamba in the Llama: Distilling and Accelerating Hybrid Models. *NeurIPS*, 2024. 2

[95] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotný. VGGT: Visual Geometry Grounded Transformer. In *CVPR*, 2025. 1, 2, 3, 4, 5, 6

[96] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3D Perception Model with Persistent State. In *CVPR*, 2025. 2, 6

[97] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D Vision Made Easy. In *CVPR*, 2024. 1, 2

[98] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. TartanAir: A Dataset to Push the Limits of Visual SLAM. In *Int. Conf. Intell. Robot. Syst.*, pages 4909–4916, 2020. ISSN: 2153-0866. 1

[99] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-Equivariant Visual Geometry Learning. *arXiv preprint arXiv:2507.13347*, 2025. 1, 2

[100] Kyle Wilson and Noah Snavely. Robust Global Translations with 1DSfM. In *ECCV*, 2014. 4

[101] Yuqi Wu, Wenzhao Zheng, Jie Zhou, and Jiwen Lu. Point3R: Streaming 3D Reconstruction with Explicit Spatial Pointer Memory. *arXiv preprint arXiv:2507.02863*, 2025. 2

[102] 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*, pages 22378–22389, 2024. 1

[103] Jianing Yang, Alexander Sax, Kevin J. Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3R: Towards 3D Reconstruction of 1000+ Images in One Forward Pass. In *CVPR*, 2025. 2

[104] Songlin Yang and Yu Zhang. Fla: A triton-based library for hardware-efficient implementations of linear attention mechanism, 2024. 2

[105] Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. *arXiv preprint arXiv:2312.06635*, 2023. 2

[106] Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing Linear Transformers with the Delta Rule over Sequence Length. In *NeurIPS*, 2024. 2, 4

[107] Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated Delta Networks: Improving Mamba2 with Delta Rule. In *ICLR*, 2025. 2, 4

[108] 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*, pages 1790–1799, 2020. 1

[109] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: A High-Fidelity Dataset of 3D Indoor Scenes. In *ICCV*, pages 12–22, 2023. 1

[110] Amir R. Zamir, Alexander Sax, William Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling Task Transfer Learning. In *CVPR*, pages 3712–3722, 2018. 1

[111] Michael Zhang, Kush Bhatia, Hermann Kumbong, and Christopher Ré. The Hedgehog & the Porcupine: Expressive Linear Attentions with Softmax Mimicry. *arXiv preprint arXiv:2402.04347*, 2024. 2

[112] Michael Zhang, Simran Arora, Rahul Chalamala, Alan Wu, Benjamin Spector, Aaryan Singhal, Krithik Ramesh, and Christopher Ré. LoLCATs: On Low-Rank Linearizing of Large Language Models. *arXiv preprint arXiv:2410.10254*, 2025. 2, 8

[113] Shangzhan Zhang, Jianyuan Wang, Yinghao Xu, Nan Xue, Christian Rupprecht, Xiaowei Zhou, Yujun Shen, and Gordon Wetzstein. FLARE: Feed-Forward Geometry, Appearance and Camera Estimation from Uncalibrated Sparse Views. In *CVPR*, 2025. 1, 2

[114] Tianyuan Zhang, Sai Bi, Yicong Hong, Kai Zhang, Fujun Luan, Songlin Yang, Kalyan Sunkavalli, William T. Freeman, and Hao Tan. Test-Time Training Done Right. *arXiv preprint arXiv:2505.23884*, 2025. 2, 5

[115] Qunjie Zhou, Sérgio Agostinho, Aljoša Ošep, and Laura Leal-Taixé. Is geometry enough for matching in visual localization? In *ECCV*, 2022. 2

[116] Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming 4D Visual Geometry Transformer. *arXiv preprint arXiv:2507.11539*, 2025. 2

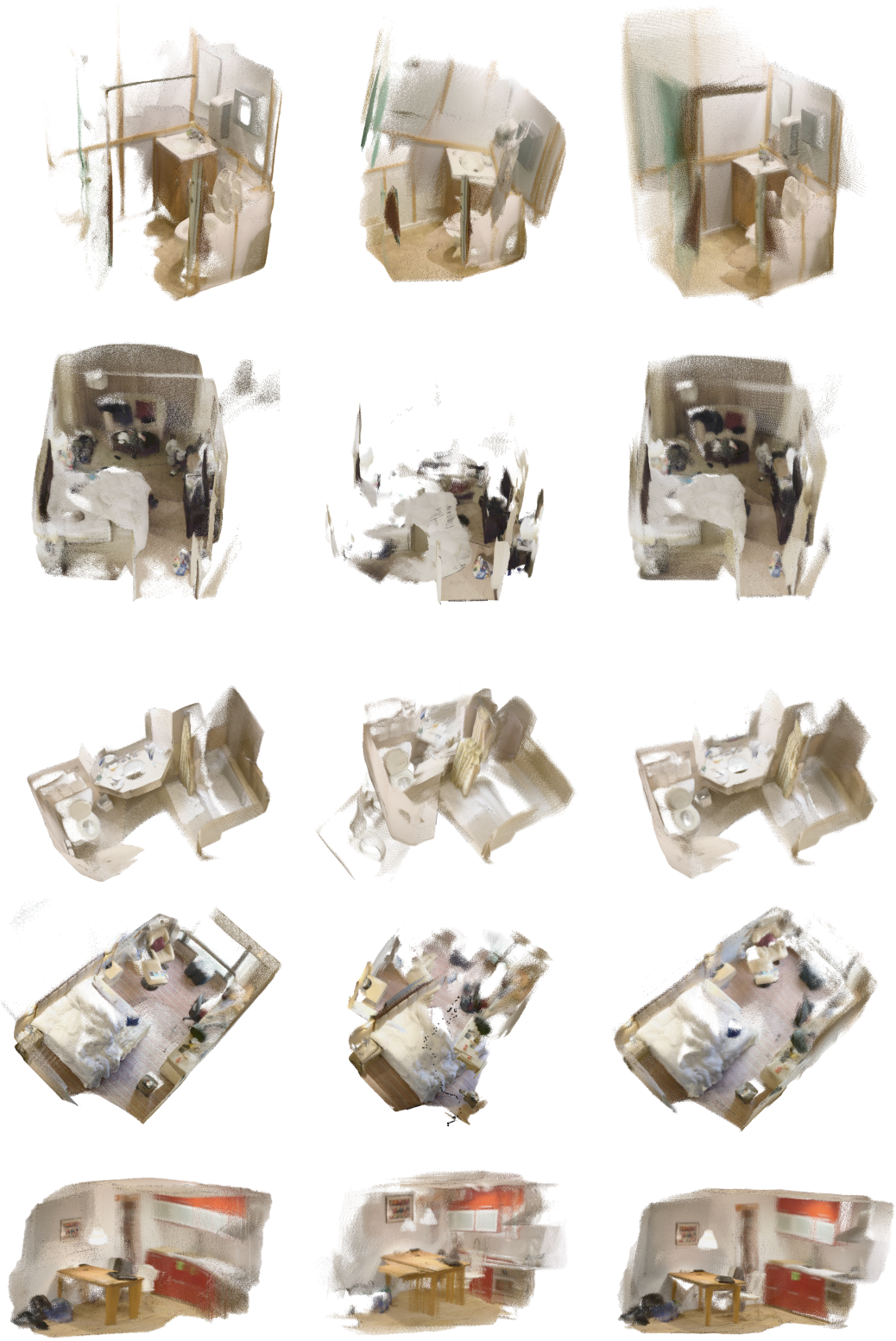
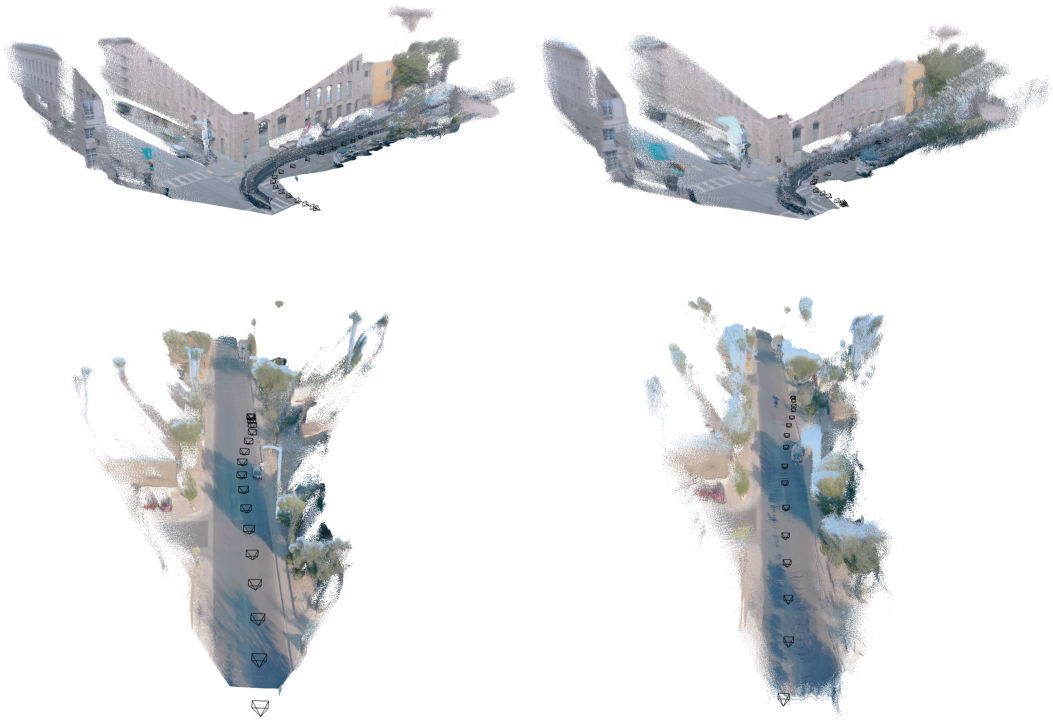


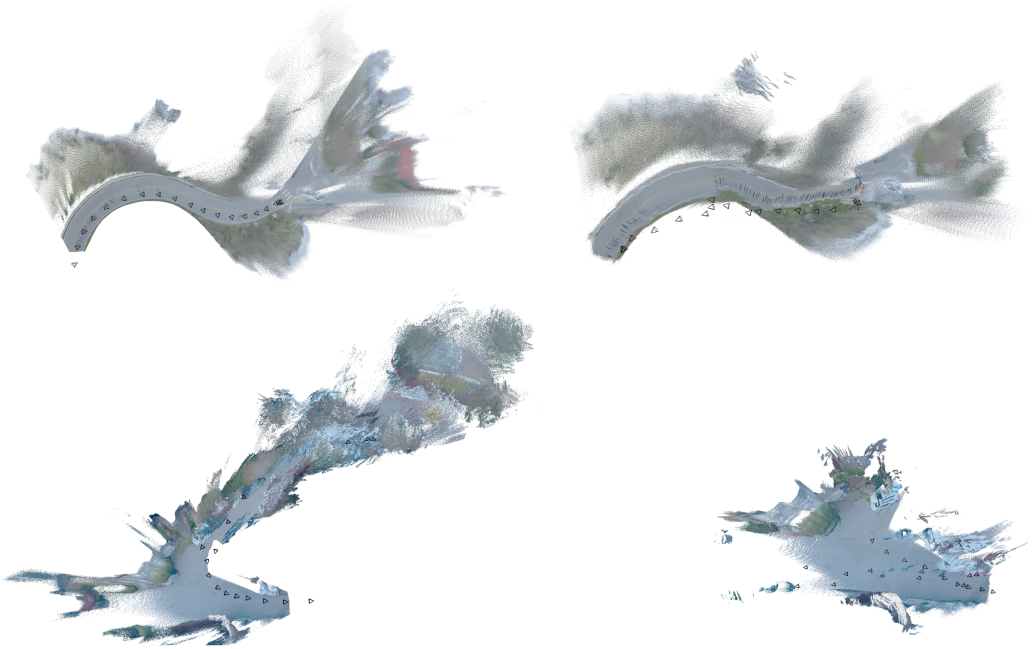
Figure 5. **Qualitative comparison.** From left to right: VGGT, TTT3R, VGG-T^3 (Ours)

Figure 6. **Visual localization examples in Wayspots and 7scenes.** Ground truth camera for query image (not used for reconstruction) shown on the left in green, predicted camera and geometry in red.

Figure 7. **In-the-wild visual localization.** We reconstruct a sequence of the KITTI dataset, then localize a tourist picture that was recorded 7 years later. Note the changes in appearance and composition of the scene.



(a) Similar reconstruction as VGGT.



(b) Failure cases.

Figure 8. Waymo sequence reconstructions comparison with VGGT.