G3AR: Graph-Guided Neural Visual Geometry for Scalable Multi-Sequence Aerial Registration

Jeng Wen Joshua Lean
National Tsing Hua University
Hsinchu, Taiwan
joshualeanjw@gmail.com

Ting-Yu Yen
National Tsing Hua University
Hsinchu, Taiwan
tingyus995@gmail.com

Wei-Fang Sun
NVIDIA AI Technology Center
Taiwan
johnsons@nvidia.com

Simon See
NVIDIA AI Technology Center
Singapore
ssee@nvidia.com

Hung-Kuo Chu
National Tsing Hua University
Hsinchu, Taiwan
hkchu@cs.nthu.edu.tw

Shih-Hsuan Hung*
National Tsing Hua University
Hsinchu, Taiwan
hungsh@cs.nthu.edu.tw

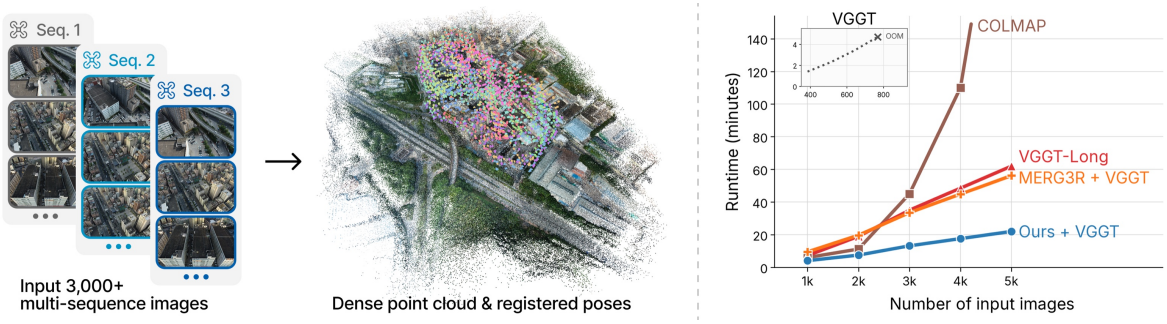

Figure 1: G3AR registers thousands of multi-sequence aerial images into a coherent point cloud with lower runtime than prior long-context pipelines.

Abstract

Full-context neural visual geometry is impractical for thousands of images, while sequence-based chunking poorly captures irregular non-local overlap in multi-sequence aerial collections. We present Graph-Guided Neural Visual Geometry for Aerial Registration (G3AR), a graph-guided framework for scalable dense neural geometry. Before local inference, G3AR builds a geometrically verified image-proximity graph that guides bounded overlapping chunks and induces a chunk graph whose maximum spanning tree defines alignment topology. Compatible backbones process chunks independently; shared-image predictions then estimate three-dimensional similarity (Sim(3)) transforms that register local cameras and geometry in a common frame. Across four real aerial scenes, G3AR improves pose error and runtime in matched VGGT- and Pi3-backed comparisons, while its DA3 variant achieves the lowest pose error among evaluated neural-geometry methods.

CCS Concepts

• **Computing methodologies** → **Computer vision problems**.

*Corresponding author.

SA Technical Communications '26, Kuala Lumpur, Malaysia

ACM ISBN 979-8-4007-2841-9/2026/12
https://doi.org/10.1145/3829339.3847824

Keywords

Neural visual geometry, aerial multi-sequence registration, image proximity graphs

ACM Reference Format:
Jeng Wen Joshua Lean, Ting-Yu Yen, Wei-Fang Sun, Simon See, Hung-Kuo Chu, and Shih-Hsuan Hung. 2026. G3AR: Graph-Guided Neural Visual Geometry for Scalable Multi-Sequence Aerial Registration. In *SIGGRAPH Asia 2026 Technical Communications (SA Technical Communications '26), December 01–04, 2026, Kuala Lumpur, Malaysia.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3829339.3847824

1 Introduction

Feed-forward neural geometry models such as VGGT [2025], Pi3 [2026], and DA3 [2025] directly predict camera poses and dense geometry. Yet full-context inference over thousands of images is impractical. Long-context methods such as VGGT-Long [2026] and MERG3R [2026] therefore process and align chunks, making results depend on chunk composition and alignment topology.

Large aerial surveys span multiple flights without a reliable global temporal order: overlap crosses strips, revisits, or vehicles, while sequence boundaries may be unrelated. Scaling therefore requires reliable non-local connections for chunk formation and alignment without exhaustive matching or full-scene optimization.

We present **G3AR**: **G**raph-**G**uided Neural Visual **G**eometry for Scalable Multi-Sequence **A**erial **R**egistration, which adapts graph partitioning to dense neural geometry. A geometrically verified proximity graph guides bounded overlapping chunks for neural

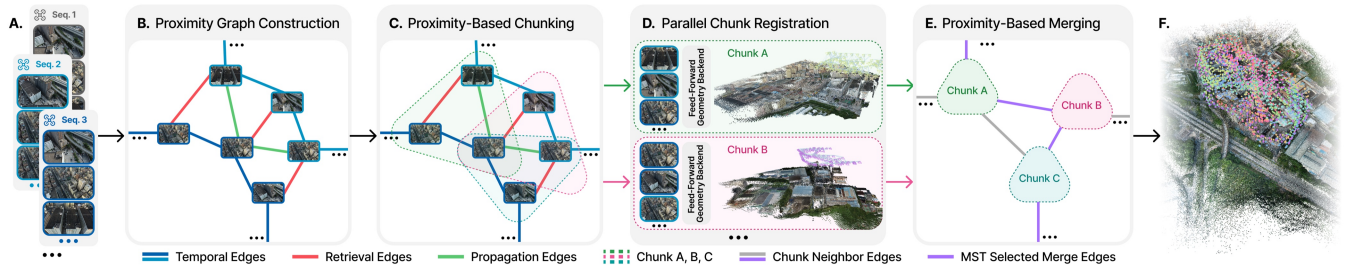

Figure 2: Overview of G3AR. (A) Multiple aerial image sequences are provided as input. (B) Verified temporal, retrieval, and propagation edges organize the images into a proximity graph. (C) Graph partitioning and boundary expansion form bounded overlapping chunks. (D) A feed-forward geometry backbone predicts chunk-local camera poses and dense geometry in parallel. (E) A maximum spanning tree of the proximity-derived chunk graph selects shared-image Sim(3) alignments. (F) Composing the transformations registers the cameras and aligns the geometry in a common frame.

inference; their chunk graph defines a maximum spanning tree for shared-image Sim(3) alignment. Compatible backbones independently predict local cameras and geometry, supporting multi-GPU inference. Assembly requires no backbone retraining, global pose-graph refinement, or post-assembly bundle adjustment. On four real aerial scenes, G3AR reduces pose error over matched VGGT- and Pi3-backed baselines, while DA3 achieves the lowest evaluated neural pose error. On the 5,621-image synthetic Small City scene from MatrixCity [2023], its VGGT variant reduces pose error, point-cloud error, and runtime relative to VGGT-Long and MERG3R. With DA3, eight L40 GPUs achieve a 2.29× mean end-to-end speedup across the four real scenes.

2 Related Work

COLMAP [2016] and GLOMAP [2024] combine verified matching, camera registration, and global optimization. GraphSfM [2020] partitions verified view graphs into overlapping clusters and merges independent reconstructions. FastMap [2026] and InstantSfM [2025] target efficiency; the supplement compares these systems and DAGSfM with G3AR. Feed-forward models VGGT [2025], Pi3 [2026], and DA3 [2025] predict cameras and dense geometry, while VGGT-Long [2026], SwiftVGGT [2026], and MERG3R [2026] scale through subset processing and alignment. G3AR adapts graph partitioning to neural geometry, using verified aerial proximity to jointly guide bounded inference chunks and inter-chunk Sim(3) alignment.

3 Method

Given images $\{I_i\}_{i=1}^{N}$ from one or more flight sequences, G3AR builds a weighted proximity graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, w)$, where $\mathcal{V} = [N]$ indexes images, $\mathcal{E}$ contains geometrically verified temporal, retrieval, and propagation edges, and w_{ij} measures verified overlap. As illustrated in Fig. 2, **(A)** multi-sequence aerial images are provided; **(B)** verified edges form a proximity graph; **(C)** each connected component is partitioned and expanded into bounded overlapping chunks $\{C_k\}$; **(D)** a compatible feed-forward geometry backbone processes chunks independently in parallel; **(E)** a maximum spanning tree of the induced chunk graph selects shared-image Sim(3) alignments; and **(F)** composing these transformations registers cameras and aligns geometry.

3.1 Proximity Graph Construction

We propose *temporal*, *retrieval*, and *propagation* edges to recover local and non-local overlap without exhaustive pairing; geometric verification filters repetitive-structure matches and determines edge retention and weight. For image i, flight metadata and UltraVPR [Chen et al. 2025] provide neighborhoods $\mathcal{N}_{\text{temp}}(i)$ and $\mathcal{N}_{\text{ret}}(i)$, respectively, yielding initial candidates $\widetilde{\mathcal{E}}_0 = \{\{i, j\} : j \in \mathcal{N}_{\text{temp}}(i) \cup \mathcal{N}_{\text{ret}}(i)\}$. Temporal candidates preserve flight continuity; retrieval identifies revisits and cross-strip or cross-sequence overlap. For each candidate $\{i, j\}$, ALIKED [2023] and LightGlue [2023] produce tentative correspondences $\mathcal{M}_{ij}$. Fundamental-matrix verification with USAC [2013] and MAGSAC [2020] yields inliers $\widehat{\mathcal{M}}_{ij}$. Let $\rho_i(\widehat{\mathcal{M}}_{ij})$ denote the fraction of grid cells in I_i containing an inlier. We define $w_{ij} = |\widehat{\mathcal{M}}_{ij}| \max\{\rho_i(\widehat{\mathcal{M}}_{ij}), \rho_j(\widehat{\mathcal{M}}_{ij})\}$, with $w_{ij} = 0$ upon verification failure. The operation Verify$(\cdot)$ retains pairs with $w_{ij} \geq \tau$, where $\tau = 200$, giving $\mathcal{E}_0 = \text{Verify}(\widetilde{\mathcal{E}}_0)$ and $\mathcal{G}_0 = (\mathcal{V}, \mathcal{E}_0, w)$. For each retained retrieval proposal $i \to j$, propagation proposes edges from i to unconnected neighbors of j in $\mathcal{G}_0$. Applying the same procedure gives $\mathcal{E} = \mathcal{E}_0 \cup \text{Verify}(\widetilde{\mathcal{E}}_{\text{prop}})$. UltraVPR only retrieves candidates; verified local geometry supports every final edge. We process $\mathcal{G} = (\mathcal{V}, \mathcal{E}, w)$ independently per connected component.

3.2 Proximity-Based Chunking

We form bounded inference chunks that preserve high-proximity neighborhoods and share images for 3D alignment. Let b_p and b_c denote the partition budget and maximum expanded-chunk size. Weighted METIS partitioning [1998] favors low-weight cuts, producing disjoint partitions $\{\mathcal{P}_k\}$ that cover each component of $\mathcal{G}$ and satisfy $|\mathcal{P}_k| \leq b_p$. Components within budget remain unchanged; those without usable internal adjacency use sequential partitioning. For each $\mathcal{P}_k$, let $\bar{\mathcal{P}}_k$ contain outside images connected to it by $\mathcal{E}$. Adding the highest-proximity images from $\bar{\mathcal{P}}_k$ forms C_k, subject to $|C_k| \leq b_c$. Shared images $O_{k\ell} = C_k \cap C_\ell$ constrain subsequent Sim(3) alignment. A compatible feed-forward geometry backbone independently predicts chunk-local camera poses and dense geometry. Backbone-independent graph construction and chunking support VGGT, Pi3, and DA3, with chunk inference distributable across GPUs.

Table 1: Average camera pose estimation on four real aerial scenes. Values are Sim(3)-aligned and averaged equally across scenes. COLMAP and GLOMAP are unranked classical references. Among neural-geometry methods, the top-3 results are highlighted as first, second, and third.

Method	ATE RMSE ↓	ATE mean ↓	ATE median ↓	Runtime ↓	Max VRAM (GiB)
COLMAP [Schönberger and Frahm 2016]	0.031	0.022	0.016	61 min 17 s	N/A
GLOMAP [Pan et al. 2024]	3.570	3.207	3.177	27 min 00 s	N/A
VGGT-Long [Deng et al. 2026]	2.355	2.084	1.994	23 min 50 s	20.5
SwiftVGGT [Lee et al. 2026]	2.370	2.073	1.870	10 min 25 s	33.5
MERG3R + VGGT [Cheng et al. 2026]	2.556	2.288	2.170	23 min 54 s	20.9
MERG3R + Pi3 [Cheng et al. 2026]	1.030	0.900	0.835	21 min 14 s	19.3
Ours + VGGT	1.855	1.529	1.317	9 min 42 s	21.7
Ours + Pi3	0.733	0.483	0.384	9 min 38 s	19.7
Ours + DA3	0.653	0.403	0.301	10 min 23 s	26.7

3.3 Proximity-Based Merging

After chunk registration, we lift image proximity to a chunk graph for lightweight assembly. Expanded chunks induce $\mathcal{G}_C = (\mathcal{K}, \mathcal{E}_C, \eta)$, where $\mathcal{K} = [N_C]$ indexes chunks and $\{k, \ell\} \in \mathcal{E}_C$ when neighboring chunks share images, i.e., $O_{k\ell} \neq \emptyset$. Weights $\eta_{k\ell}$ combine proximity evidence and shared-image overlap. A maximum spanning tree is extracted independently per connected component, prioritizing strong connections without sequential ordering. We define $\eta_{k\ell} = |O_{k\ell}| \max w_{ij}$ over cross-partition edges. For each selected edge $\{k, \ell\}$, dense predictions of images in $O_{k\ell}$ provide corresponding 3D points in both chunk frames. Following VGGT-Long [2026], G3AR uses iteratively reweighted least squares to estimate a robust pairwise Sim(3) transformation, recovering relative scale, rotation, and translation. For each tree, we select a reference chunk and compose pairwise Sim(3) transformations along unique tree paths, placing all reachable cameras and geometry in a common frame. This tree-based assembly requires neither global pose-graph Levenberg-Marquardt refinement nor post-merge bundle adjustment.

4 Experiments

We evaluate Building and Rubble from Mill19 [2022] and Residence and Sci-Art from UrbanScene3D [2022] (1,657–2,998 images), plus the 5,621-image synthetic Small City scene from MatrixCity [2023], with ground-truth poses and a dense reference point cloud. We report camera coverage, Sim(3)-aligned ATE, end-to-end runtime, peak GPU memory, and Chamfer-L1 where available, using one NVIDIA L40 GPU unless noted. For controlled comparison, VGGT-Long, MERG3R, and G3AR use at most 95 images per chunk with a 35-image overlap target ($b_p = 75$, $b_c = 95$), retaining their own chunking and alignment strategies. The supplement provides broader classical SfM comparisons, scaling details, and ablations of temporal-order dependence, alignment topology, and global optimization.

Quantitative Results. Full-context runs of VGGT [2025], Pi3 [2026], DA3 [2025], FastVGGT [2026], and LiteVGGT [2026] exceed 48 GiB and are omitted. G3AR reduces ATE and runtime relative to VGGT-Long and MERG3R with VGGT, and to MERG3R with Pi3. All three variants maintain full camera coverage; Ours + DA3 achieves the lowest ATE and Ours + Pi3 the shortest runtime among evaluated neural methods (Table 1). On Small City [2023], Ours + VGGT achieves the lowest ATE, Chamfer-L1, and runtime among the compared neural pipelines (Table 2).

Table 2: Small City results with VGGT local geometry. Errors are Sim(3)-aligned in dataset units.

Method	ATE RMSE ↓	Runtime ↓	Chamfer-L1 ↓
VGGT-Long	6.6277	44 min 55 s	2.7457
MERG3R + VGGT	5.2523	44 min 3 s	3.1699
Ours + VGGT	**4.7639**	**24 min 8 s**	**1.3408**

Table 3: DA3 chunk-image selection ablation, averaged equally over four real scenes.

Variant	ATE RMSE ↓
Temporal edges only	2.625
Temporal + retrieval edges	0.819
Temporal + retrieval + propagation edges	0.653

Qualitative Results. On Building and Residence, graph-guided variants preserve coherent global layouts and broader extents than the displayed long-context baselines, while local geometry remains more diffuse than COLMAP (Fig. 3).

Multi-GPU Scaling. Across four real scenes with DA3, eight warmed L40 GPUs achieve 7.11× mean chunk speedup at 88.9% efficiency and 2.29× mean end-to-end speedup (Fig. 4).

Ablations. Chunk-selection ablations with DA3 across four real scenes show that retrieval reduces ATE RMSE over temporal edges alone, with further gains from propagation (Table 3), supporting verified non-local edges for chunk selection.

5 Conclusion

G3AR uses graph-guided partitioning, shared-image overlap, and inter-chunk Sim(3) alignment to scale dense neural geometry for multi-sequence aerial registration. Across four real scenes, all three backbones retain full coverage and improve pose error over matched long-context counterparts; eight-GPU chunk inference achieves 7.11× mean speedup (2.29× end-to-end). Future work will improve graph robustness beyond aerial scenes and support incremental reconstruction.

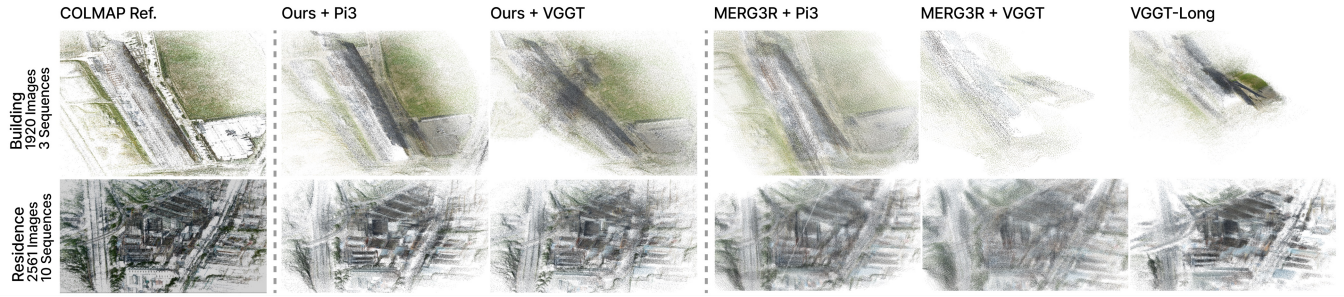

Figure 3: Building and Residence point clouds under matched local inference budgets. Graph-guided variants retain broader roof, facade, and ground extents than the displayed long-context baselines.

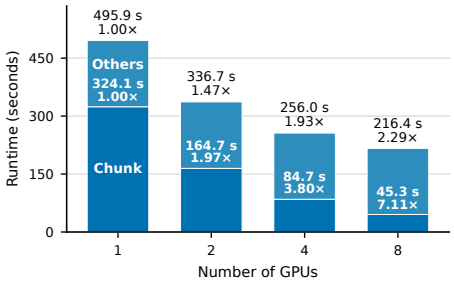

Figure 4: Scaling over four real scenes. Labels show mean runtime and mean per-scene speedup relative to one GPU.

Acknowledgments

We thank the anonymous reviewers. This work was supported by Taiwan's National Science and Technology Council grants 114-2221-E-007-114-MY3, 113-2221-E-007-102-MY3, 114-2221-E-007-115-MY3, and 115-2634-F-007-003. We also thank NVIDIA Corporation and NVIDIA AI Technology Center (NVAITC) for providing access to the Taipei-1 supercomputer.

References

Daniel Barath, Jana Noskova, Maksym Ivashechkin, and Jiří Matas. 2020. MAGSAC++, a Fast, Reliable and Accurate Robust Estimator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1304–1312.

Chao Chen, Chunyu Li, Mengfan He, Jun Wang, Fei Xing, and Ziyang Meng. 2025. UltraVPR: Unsupervised Lightweight Rotation-Invariant Aerial Visual Place Recognition. *IEEE Robotics and Automation Letters* 10, 9 (2025), 9096–9103. doi:10.1109/LRA.2025.3592075

Yu Chen, Shuhan Shen, Yisong Chen, and Guoping Wang. 2020. Graph-based Parallel Large Scale Structure from Motion. *Pattern Recognition* 107 (2020), 107537. doi:10.1016/j.patcog.2020.107537

Leo Kaixuan Cheng, Abdus Shaikh, Ruofan Liang, Zhijie Wu, Yushi Guan, and Nandita Vijaykumar. 2026. MERG3R: A Divide-and-Conquer Approach to Large-Scale Neural Visual Geometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 28969–28978.

Kai Deng, Zexin Ti, Jiawei Xu, Jian Yang, and Jin Xie. 2026. VGGT-Long: Chunk it, Loop it, Align it – Pushing VGGT's Limits on Kilometer-scale Long RGB Sequences. In *Proceedings of the IEEE International Conference on Robotics and Automation*.

George Karypis and Vipin Kumar. 1998. A Fast and High Quality Multilevel Scheme for Partitioning Irregular Graphs. *SIAM Journal on Scientific Computing* 20, 1 (1998), 359–392. doi:10.1137/S1064827595287997

Jungho Lee, Minhyeok Lee, Sunghun Yang, Minseok Kang, and Sangyoun Lee. 2026. SwiftVGGT: A Scalable Visual Geometry Grounded Transformer for Large-Scale Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*. 447–456.

Jiahao Li, Haochen Wang, Muhammad Zubair Irshad, Igor Vasiljevic, Matthew R. Walter, Vitor Campagnolo Guizilini, and Greg Shakhnarovich. 2026. FastMap: Revisiting Structure from Motion through First-Order Optimization. In *Proceedings of the International Conference on 3D Vision*. 29–39. doi:10.1109/3DV69130.2026.00010

Yixuan Li, Lihan Jiang, Linning Xu, Yuanbo Xiangli, Zhenzhi Wang, Dahua Lin, and Bo Dai. 2023. MatrixCity: A Large-scale City Dataset for City-scale Neural Rendering and Beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3205–3215.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. 2025. Depth Anything 3: Recovering the Visual Space from Any Views. arXiv preprint arXiv:2511.10647.

Liqiang Lin, Yilin Liu, Yue Hu, Xingguang Yan, Ke Xie, and Hui Huang. 2022. Capturing, Reconstructing, and Simulating: The UrbanScene3D Dataset. In *Proceedings of the European Conference on Computer Vision*. 93–109. doi:10.1007/978-3-031-20074-8_6

Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. 2023. LightGlue: Local Feature Matching at Light Speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 17627–17638.

Linfei Pan, Daniel Barath, Marc Pollefeys, and Johannes L. Schönberger. 2024. Global Structure-from-Motion Revisited. In *Proceedings of the European Conference on Computer Vision*. 58–77. doi:10.1007/978-3-031-73661-2_4

Rahul Raguram, Ondřej Chum, Marc Pollefeys, Jiří Matas, and Jan-Michael Frahm. 2013. USAC: A Universal Framework for Random Sample Consensus. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 8 (2013), 2022–2038. doi:10.1109/TPAMI.2012.257

Johannes L. Schönberger and Jan-Michael Frahm. 2016. Structure-from-Motion Revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4104–4113.

You Shen, Zhipeng Zhang, Yansong Qu, Xiawu Zheng, Jiayi Ji, Shengchuan Zhang, and Liujuan Cao. 2026. FastVGGT: Fast Visual Geometry Transformer. In *International Conference on Learning Representations*.

Zhijian Shu, Cheng Lin, Tao Xie, Wei Yin, Ben Li, Zhiyuan Pu, Weize Li, Yao Yao, Xun Cao, Xiaoyang Guo, and Xiao-Xiao Long. 2026. LiteVGGT: Boosting Vanilla VGGT via Geometry-aware Cached Token Merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 36422–36432.

Haithem Turki, Deva Ramanan, and Mahadev Satyanarayanan. 2022. Mega-NeRF: Scalable Construction of Large-Scale NeRFs for Virtual Fly-Throughs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12922–12931.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. 2025. VGGT: Visual Geometry Grounded Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5294–5306.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. 2026. π^3: Permutation-Equivariant Visual Geometry Learning. In *International Conference on Learning Representations*.

Xiaoming Zhao, Xingming Wu, Weihai Chen, Peter C. Y. Chen, Qingsong Xu, and Zhengguo Li. 2023. ALIKED: A Lighter Keypoint and Descriptor Extraction Network via Deformable Transformation. *IEEE Transactions on Instrumentation and Measurement* 72 (2023), 1–16. doi:10.1109/TIM.2023.3271000

Jiankun Zhong, Zitong Zhan, Quankai Gao, Ziyu Chen, Haozhe Lou, Jiageng Mao, Ulrich Neumann, Chen Wang, and Yue Wang. 2025. InstantSfM: Towards GPU-Native SfM for the Deep Learning Era. arXiv preprint arXiv:2510.13310.