The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-1/W6-2025
3rd International Workshop on Evaluation and BENCHmarking of Sensors, Systems and
GEOspatial Data in Photogrammetry and Remote Sensing (GEOBENCH), 20–21 November 2025, Wroclaw, Poland

Exploring modern end-to-end AI-based multi-view 3D reconstruction

June Moh Goo[1], Zichao Zeng[1], Luca Morelli[2], Fabio Remondino[2], Jan Boehm[1]

[1] Department of Civil, Environmental and Geomatic Engineering, University College London, UK
Email: <june.goo.21><zichao.zeng.21><j.boehm>@ucl.ac.uk
[2] 3D Optical Metrology (3DOM) unit, Bruno Kessler Foundation (FBK), Trento, Italy
Email: <lmorelli><remondino>@fbk.eu

Keywords: Image Orientation, Bundle Adjustment, Foundation Model, Deep Learning, 3D Reconstruction.

Abstract
Deriving accurate 3D geometry from multi-view 2D imagery remains a fundamental problem in photogrammetry and computer vision. Conventional pipelines, comprising feature extraction, image matching, bundle adjustment and dense reconstruction, are grounded in well-established geometric principles but remain sensitive to complex scenarios such as significant illumination variability, deficiency in texture and high variability in viewing angles. Recent deep learning developments have triggered a paradigm shift, reformulating multi-view 3D reconstruction as a data-driven, end-to-end optimization problem. Neural architectures now jointly learn feature representations, correspondence estimation and geometric reasoning, supported by large-scale training datasets, high-performance GPU computation, transformer networks and differentiable rendering frameworks. This study methodically examines the transition from traditional photogrammetric approaches to end-to-end AI-based reconstruction pipelines. Using benchmark geomatic datasets, we quantitatively evaluate the performance of two recent and representative end-to-end deep learning methods compared to classical photogrammetry. Results highlight performances of AI-driven approaches in 3D reconstructions and their limits for in large-scale, metric-oriented mapping and modeling applications.

1. Introduction

The fundamental challenge of deriving 3D geometry from multi-view 2D images has attracted photogrammetry and computer vision people for decades (Pollefeys et al., 2004; Remondino and El-Hakim, 2006; Agarwal et al., 2009).

Nowadays, 3D scene reconstruction is becoming increasingly important, with applications ranging from autonomous driving to urban planning, digital twins and heritage documentation. What began as a purely geometric problem, traditionally solved through classical photogrammetric principles and structure-from-motion (SfM) techniques (Schonberger and Frahm, 2016; Remondino et al., 2017), is undergoing a revolutionary transformation with the advent of deep learning methods for tie point extraction, outlier removal, dense image matching or end-to-end 3D reconstructions (Yao et al., 2021; Stathopoulou and Remondino, 2023; Morelli et al., 2024; Keetha et al., 2025; Perda et al., 2025).

Traditional approaches, whilst mathematically elegant and interpretable, often struggle with challenging scenarios (e.g. textureless surfaces, repetitive patterns, dynamic lighting conditions, etc.) or large datasets. Moreover, the general sequential processing nature of traditional pipelines, from feature extraction, matching and bundle adjustment to dense reconstruction, creates error propagation pathways where failures in early stages affect the entire process.

The recent emergence of deep learning has fundamentally changed also 3D processing workflows. Modern AI-based approaches treat multi-view reconstruction as an end-to-end optimization problem, based on implicit representations, leveraging neural networks to learn robust feature representations, correspondence estimation and 3D data estimation directly from data (Wang et al., 2024a; Wang et al, 2024b; Yang et al., 2025). Additionally, differentiable rendering pipelines make it possible to optimize 3D predictions directly from image-level supervision, reducing the dependence on ground-truth geometry (Navaneet et al., 2019; Gao and Qi, 2024). This transformation has been facilitated by several key technological advances: the availability of large-scale synthetic and real-world image datasets, increased computational power (GPU) and architectural innovations including attention mechanisms, transformer networks and differentiable rendering techniques. Despite these achievements, challenges remain. Metric accuracy, scalability and generalization across domains continue to limit the adoption of purely AI-based methods in professional 3D mapping and photogrammetry. Hybrid frameworks that combine geometric constraints with learned priors are emerging as a promising direction, offering the interpretability of classical methods with the adaptability and robustness of deep networks (Yin et al., 2023; Mu et al., 2023).

1.1 Aim of the work

This work examines the current state of end-to-end AI-based multi-view 3D reconstruction, analyzing the transition from traditional photogrammetric approaches to modern deep learning architectures (Table 1). Using common geomatic datasets, the aim of the paper is two-fold:

- to investigate how end-to-end AI-based methods perform with respect to classical photogrammetry;
- to show how the convergence (or replacement?) of classical geometric principles with modern deep learning architecture still faces some critical issues in terms of accuracy and scalable 3D reconstruction results.

2. End-to-end 3D reconstruction methods

Today's AI-based multi-view 3D reconstruction represents a paradigm shift from hand-crafted algorithms to end-to-end learned systems that can automatically extract, match and triangulate features while simultaneously estimating camera poses and scene geometry. Proposed methods initially solved for feature extraction (Yao et al., 2021) or camera poses (Kendall et al., 2015; Wang et al., 2024a) and MVS (Wang et al., 2022) till complete end-to-end 3D scene reconstruction: 3D-RETR (Shi et al., 2021), DUSt3R (Wang et al., 2024b), MASt3R (Leroy et al., 2024), PF-LRM (Wang et al., 2024c), VGGT (Wang et al., 2025a), Spann3R (Wang and Agapito, 2025), MapAnything

	Photogrammetry (Conventional Geometric Methods)	***AI-based (End-to-End methods)***
Approach	Use explicit geometric models (pair-wise epipolar geometry, collinearity principle, triangulation / bundle adjustment)	Data-drive, learning a direct mapping from images to 3D data using deep neural networks (e.g., transformers, implicit neural fields).
Processing pipeline	Sequential for feature detection, matching, camera pose estimation and dense 3D reconstruction	Generally end-to-end / unified for feature extraction, matching, depth estimation, fusion, 3D reconstruction
Input data	Images and, ideally, intrinsic parameters or exterior orientation approximations to speed up calculations	Images
Calibration	If not available, derived in self-calibration	Inferred
Feature extraction	Handcrafted or learning-based methods	Learned methods
Image matching	Explicit geometric matching using descriptors or learning-based approaches	Implicit or learned correlation
Depth / 3D reconstruction	Least-squares triangulation/bundle and multi-view stereo algorithms	Learned depth prediction or neural implicit fields
Robustness	Sensitive to texture, lighting and occlusions	Learn priors for missing data and low-texture areas
Sparse views	Could fail due to lack of well distributed tie points	Efficient
Automation	Require pipeline configuration and parameter tuning	Fully automated with minimal manual tuning
Generalization	Scene-agnostic	Weak, depending on training data
Explainability	Clear statistics and metrics	No interpretability and physical rigor
Photorealism	Struggle with view-dependent effects with a rendered view quality limited by texture mapping	Capture complex, view-dependent effects like reflections, transparency and soft shadows, resulting in high-fidelity novel views
Processing	Time-consuming as images are processed in full-res	Very fast as images are downscaled due to training needs

Table 1. Main differences of conventional geometric and end-to-end methods for image-based 3D reconstruction purposes.

(Keetha et al., 2025), FAST3R (Yang et al., 2025), p3 (Wang et al., 2025b). These methods differ in their core architectures, the 3D representation they produce and their primary focus (e.g. speed, accuracy or scalability).
In this work, we focus on two recent methods that demonstrated remarkable results: we briefly introduce them in the following sections before running a quantitative evaluation in Sections 3 and 4.

2.1 VGGT

Visual Geometry Grounded Transformer (VGGT) is a feed-forward neural network that performs 3D reconstruction from one, a few or even hundreds of input views of a scene (Wang et al., 2025a). VGGT substantial departs from DUSt3R (Wang et al., 2024b), MASt3R (Leroy et al., 2024) or VGGSfM (Wang et al., 2024a) - which still require costly iterative post-optimization processes - and predicts a full set of 3D attributes, including camera extrinsic, depth maps, 3D point maps and point tracks. It does so quickly in a single forward pass, out-performing optimization-based alternatives even without further processing.

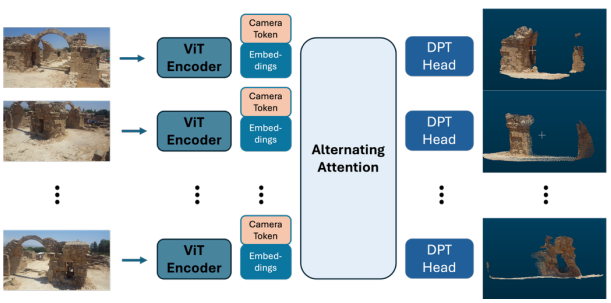

Figure 1. Model architecture of VGGT with examples on the Cyprus dataset.

As illustrated in Figure 1, VGGT consists of three main components. Each input image is first processed by a Vision Transformer (ViT) encoder, which patchifies the image into tokens and generates a set of embeddings augmented with a dedicated camera token for camera parameter prediction. These embeddings are then passed into an Alternating Attention module, which alternates between frame-wise self-attention and global self-attention layers. This alternating design enables the model to balance local feature extraction within each frame and global geometric consistency across all views. Finally, a set of Dense Prediction Transformer (DPT) heads decode the aggregated features to produce dense 3D outputs, including depth maps, 3D point maps and 3D point tracks for each input frame. The entire pipeline operates in a fully feed-forward manner, requiring no explicit geometric optimization or feature matching at inference time, leading to significant efficiency and scalability gains.
VGGT is based on a large transformer, with no particular 3D or other inductive biases (except for alternating between frame-wise and global attention) but trained on a large number of publicly available datasets with 3D annotations including CO3D (Reizenstein et al., 2021), BlendMVS (Yao et al., 2020a), DL3DV (Ling et al., 2024), MegaDepth (Yao et al., 2020b), Kubric (Greff et al., 2022), WildRGB (Xia et al., 2024), ScanNet (Dai et al., 2017), HyperSim (Roberts et al., 2021), Mapillary (Antequera et al., 2020), Habitat (Savva et al., 2019), Replica (Straub et al., 2019), MVS-Synth (Huang et al., 2018), PointOdyssey (Zheng et al., 2023), Virtual KITTI (Cabon et al., 2020), Aria Synthetic Environments (Pan et al., 2023), Aria Digital Twins (Pan et al., 2023) and an artist-created synthetic dataset similar to Objaverse (Deitke et al., 2023).
Thanks to these design choices and its feed-forward efficiency, VGGT is declared to be particularly well suited for large-scale multi-view reconstruction, real-time scene understanding, SLAM initialization and as a versatile 3D feature backbone for downstream tasks such as dense point tracking and feed-forward novel view synthesis.

2.2 Fast3R

Fast3R - Fast 3D Reconstruction via Feed-Forward Transformers (Yang et al., 2025) is a feed-forward multi-view generalization of DUSt3R (Wang et al., 2024b) that reconstructs scenes from hundreds to over a thousand unordered, unposed images in a single forward pass. It jointly reasons over all available images

using all-to-all self-attention and predicts per-view local and global point maps alongside confidence maps, from which camera poses and depth can be derived. Unlike DUSt3R or MASt3R which remain pairwise and depend on costly per-scene global alignment, or Spann3R (Wang and Agapito, 2025), which processes frames sequentially, Fast3R overcomes these limitations by processing all images simultaneously, reducing error accumulation (Yao et al., 2020b; Li and Snavely, 2018). Architecturally, Fast3R employs a standard ViT-family fusion transformer without explicit 3D-specific inductive biases and is trained on large 3D annotated datasets: CO3D (Reizenstein et al., 2021), ScanNet++ (Yeshwanth et al., 2023), ARKitScenes (Dehghan et al., 2021), Habitat (Savva et al., 2019), BlendedMVS (Yao et al., 2020b) and MegaDepth (Li and Snavely, 2018). Image-index positional embeddings with positional interpolation enable training on some 20 views but inference on more than 1000 (Yang et al., 2025). In reported settings, Fast3R achieves strong pose and reconstruction accuracy while operating at approximately 251 FPS, achieving these speed and scale benefits without the need for per-scene global alignment optimization.

Building on these advantages, Fast3R is declared to be particularly well-suited for large-scale 3D scene reconstruction in practical settings. It can process some thousands of unposed images simultaneously, allowing efficient reconstruction of detailed indoor or urban environments without explicit alignment or optimization. Compared to traditional SfM pipelines that rely on iterative matching and bundle adjustment, or pairwise learning-based models such as DUSt3R, Fast3R achieves consistent, large-scale reconstruction in real time (Zhang et al., 2025). These strengths make it valuable for robotic navigation or AR/VR environment generation.

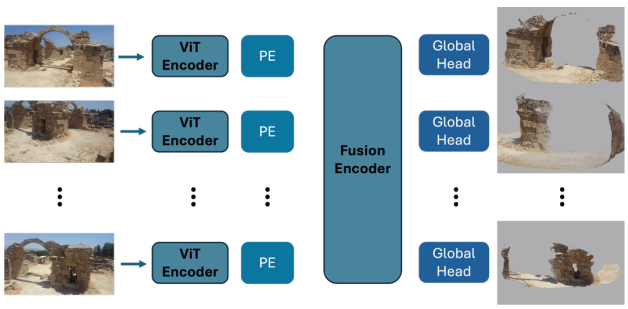

Figure 2. Model architecture of Fast3R with examples on the Cyprus dataset.

3. Datasets and evaluation

3.1 Datasets

To investigate the interplay between traditional photogrammetric pipelines (COLMAP) and emerging end-to-end AI-based multi-view 3D reconstruction methods (VGGT and Fast3R), three representative datasets included in NeRFBK[1] (Yan et al., 2023) are considered (Table 2):

- The Cyprus dataset (178 images, 15,753 pairs) originates from a terrestrial photogrammetric campaign of the Saranta Kolones archeological site in Paphos (Cyprus). The acquisitions aimed at comparing smartphone and DSLR images for photogrammetric reconstructions. In this work, the ground-truth camera poses are computed using full image resolution on the smartphone images, incorporating ten ground control points (GCPs) surveyed with a total station.
- The UAV dataset (224 images, 24,976 pairs) is also part of *UseGeo* (Nex et al., 2024), an ISPRS initiative providing multi-sensor data acquired from UAV platforms. One image block is processed at full resolution to derive ground-truth camera poses. Ten GCPs, manually identified from the available LiDAR point cloud, are used to enhance the georeferencing accuracy of the photogrammetric block.
- The Dortmund dataset (59 images, 1,711 pairs) consists of multi-sensor imagery acquired from an aerial platform (Nex et al., 2015) over the city of Dortmund (Germany). Ground truth data are generated including ten GCPs manually extracted from the corresponding aerial LiDAR point cloud.

Datasets	**Cyprus**	**UAV**	**Dortmund**
view			
platform	terrestrial	drone	airborne
# img.	178	224	59
resolution	3840x2160 px	7952x5304 px	8176x6132 px

Table 2. Summary of employed datasets[1].

3.2 Metrics

The mean Average Accuracy (mAA) on camera poses is adopted as the primary evaluation metric, following a standard evaluation protocol widely used in computer vision for the evaluation of SfM algorithms[2]. This metric allows a fair comparison between traditional geometry-based methods (in our case COLMAP) and recent AI-based foundation models (Fast3R and VGGT), which directly learn geometric relations via end-to-end neural inference of poses and depths.

The mAA metric consists of two complementary components that capture different aspects of pose estimation error, both ranging from 0 to 1, with 1 indicating the maximum correctness:
- mAA_q (mean Average Angular Accuracy) - It reflects rotational (angular) accuracy, evaluated over thresholds ranging from 1° to 30°. It measures the percentage of estimated camera orientations whose angular error is below a given threshold. For each possible camera pair, their relative rotation is computed and compared with the ground truth relative rotation. Formally, the angular error between an estimated rotation R_i and the ground-truth rotation R_i^* is computed as:

$$e_{R,i} = \text{across}\left(\frac{\text{trace}(R_i^T R_i^*)-1}{2}\right) \quad \text{(Eq. 1)}$$

The mAA_q is then obtained by integrating the accuracy over the range of angular thresholds:

$$mAA_q = \frac{1}{|\Theta|}\sum_{\theta\in\Theta}\frac{1}{N}\sum_{i=1}^{N} 1(e_{R,i} < \theta) \quad \text{(Eq. 2)}$$

where $\Theta = \{1°, 2°, \ldots, 30°\}$, N is the number of camera pairs, and $1(\cdot)$ is the indicator function.

- mAA_t (mean Average Translational Accuracy) - It reflects translational (positional) accuracy, evaluated over thresholds ranging from 0.01 m to 5 m. It measures the percentage of estimated camera centers whose positional deviation from the ground truth is below a threshold. For each image pair the translation error is defined in the local reference system of one of the two cameras taken as reference. The translational error for each pose is defined as:

$$e_{t,i} = \|t_i - t_i^*\|_2 \quad \text{(Eq. 3)}$$

and the mAA_t is computed analogously as

[1] https://github.com/3DOM-FBK/NeRFBK

[2] https://image-matching-workshop.github.io/

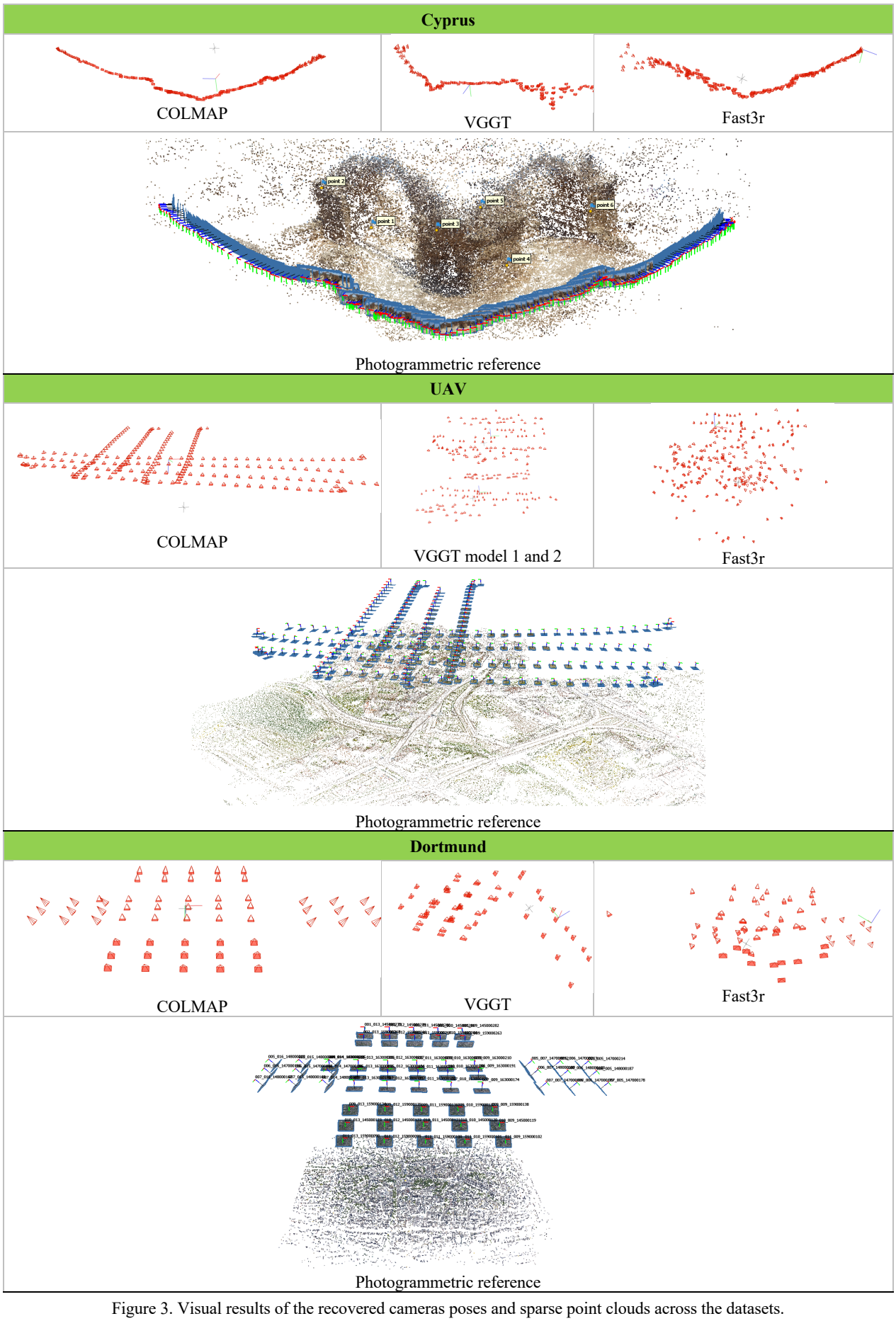

Figure 3. Visual results of the recovered cameras poses and sparse point clouds across the datasets.

$$mAA_t = \frac{1}{|T|}\sum_{\tau \in T}\frac{1}{N}\sum_{i=1}^{N} 1(e_{t,i} < \tau) \quad \text{(Eq. 4)}$$

where $T = \{0.01\ m, 0.02\ m, \ldots, 5\ m\}$.

4. Results

For the considered datasets (Section 3.1) and metrics (Section 3.2), VGGT and Fast3R results, together with the COLMAP conventional approach, are reported in Table 2 and Figure 3 in relation to the photogrammetric reference (ground truth) derived with Agisoft Metashape (Agisoft LLC, 2025).Figure 3 presents the poses estimated by the different approaches, allowing a direct qualitative comparison with the reference one. It is immediately clear that only COLMAP produces results consistent with the ground truth, whereas VGGT and Fast3D struggle to process UAV and aerial datasets, yielding highly inconsistent camera orientations.

Considering the metrics, for the Cyprus dataset, COLMAP performs closely to the reference, Fast3R partially captures the 3D structure while VGGT exhibits strong deviations. For the UAV dataset, COLMAP derived highly consistent results with the reference, VGGT generates two unstable models and Fast3R fails entirely. For the Dortmund dataset, COLMAP produces highly consistent results, whereas Fast3R and VGGT significantly struggle.

Dataset	Method	$mAA_q \uparrow$	$mAA_t \uparrow$
Cyprus	COLMAP+RootSIFT	0.889	0.449
	VGGT	0.109	0.118
	Fast3R	0.523	0.282
UAV	COLMAP+RootSIFT	1.000	0.793
	VGGT	0.267	0.098
	Fast3R	0.124	0.095
Dortmund	COLMAP+RootSIFT	0.927	0.209
	VGGT *	0.15/0.15	0.001/0.001
	Fast3R	0.025	0.001

Table 2: Quantitative comparison of COLMAP, VGGT and Fast3R across datasets. *Due to memory limitations, VGGT divides the UAV dataset into two batches, corresponding to two (3D) models.

When evaluating the results, an additional factor must be considered: image resolution during the processing phase. Both VGGT and Fast3r are computationally intensive neural networks and trained on small image sizes. Consequently, the original input images are internally automatically resized to a certain resolution suitable for the trained method (Table 3). This resizing substantially reduces the radiometric content of the images, which in turn contributes to the lower final accuracy of the 3D results. Visual results of final 3D reconstructions are reported in Figure 4 with close-up views for the Cyprus dataset in Figure 5.

Datasets	VGGT	Fast3R
Cyprus (3840x2160 px)	518x294 px	512x288 px
UAV (7952x5304 px)	518x350 px	512x336 px
Dortmund (8176x6132 px)	518x518 px	512x384 px

Table 3. Image downscaling in end-to-end methods.

5. Conclusions

This study presented a quantitative evaluation of recent end-to-end deep learning-based multi-view 3D reconstruction methods in comparison with conventional photogrammetric pipelines. The analysis employed representative datasets from the geomatics community, encompassing terrestrial, UAV and aerial image acquisitions, to assess performance across diverse acquisition geometries and scales.

Although the rapid development of AI-driven 3D reconstruction frameworks marks a significant step forward in automation and adaptability, the experimental results reveal that their current performance still limits their deployment in large-scale, metric-oriented mapping and modeling applications. In particular, camera pose estimation exhibited notable deviations from high-precision reference data, propagating errors in subsequent depth and geometry estimation.

Overall, the findings suggest that, for the used datasets, deep learning architectures have not yet reached the level of geometric accuracy, consistency and scalability required to replace traditional photogrammetric approaches in professional-grade 3D mapping applications. Instead, a hybrid or complementary integration between learning-based and geometry-based methods appears to be the most promising future direction, leveraging the robustness and interpretability of geometric models with the adaptability and efficiency of modern AI systems.

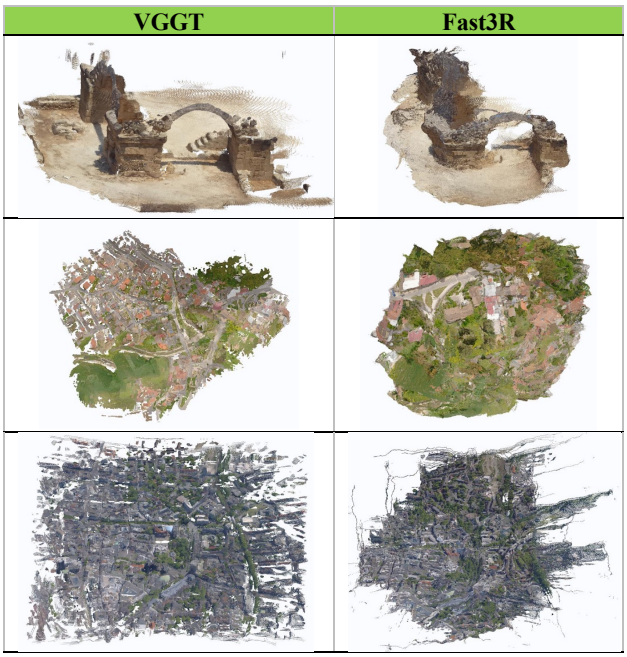

Figure 4. Visuals of the final 3D reconstruction of the end-to-end methods (Cyprus: top; UAV: centre; Dortmund: bottom).

Acknowledgments

June Moh Goo and Zichao Zeng are supported by the Engineering and Physical Sciences Research Council through an industrial CASE studentship with Ordnance Survey (Grant number EP/X524840/1 and EP/W522077/1).

References

Agarwal, S., Snavely, N., Seitz, S., Szeliski, R., 2009. Bundle adjustment in the large. Proc. *ECCV*, pp. 29-42

Agisoft LLC, 2025. Agisoft Metashape professional. https://www.agisoft.com/. Version 2.2.

Antequera, M. L., Gargallo, P., Hofinger, M., Bulo, S. R., Kuang, Y., Kontschieder, P., 2020. Mapillary planet-scale depth dataset.

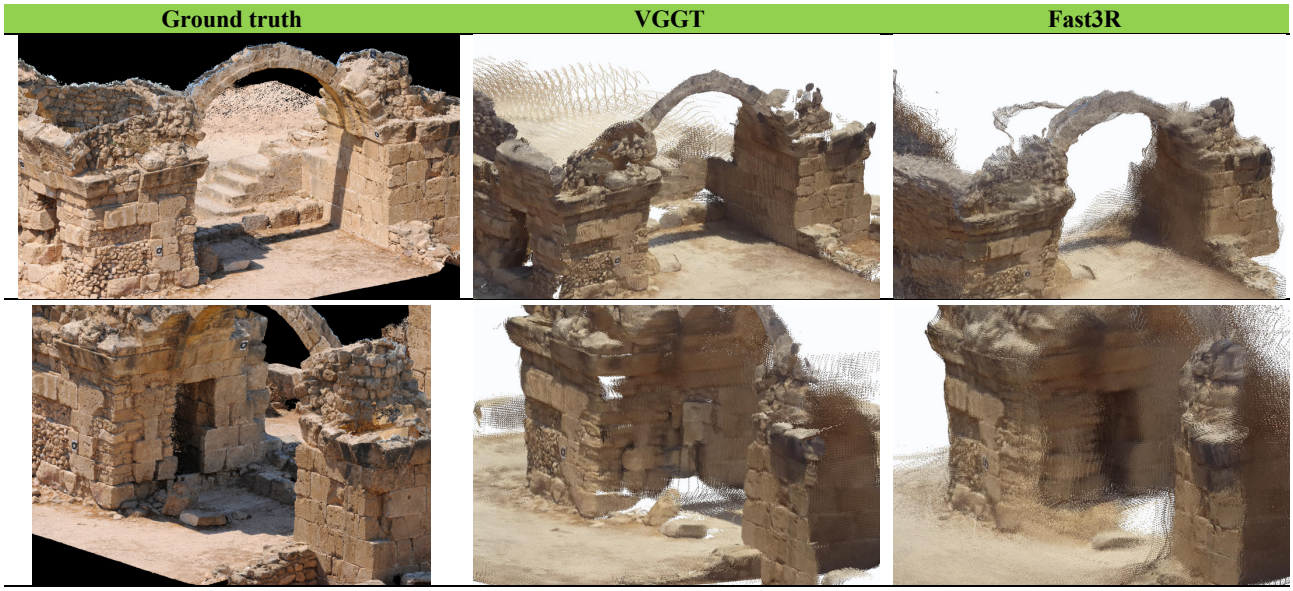

Figure 5. Close views of the final end-to-end 3D reconstruction of the Cyprus dataset with respect to the ground truth point cloud derived by classical photogrammetry.

Proc. *ECCV*, 589–604.

Cabon, Y., Murray, N., Humenberger, M., 2020. Virtual kitti 2. *arXiv preprint arXiv:*2001.10773.

Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T., Nießner, M., 2017. Scannet: Richly-annotated 3D reconstructions of indoor scenes. Proc. *CVPR*, 5828–5839.

Dehghan, A., Baruch, G., Chen, Z., Feigin, Y., Fu, P., Gebauer, T., Kurz, D., Dimry, T., Joffe, B., Schwartz, A. et al., 2021. ARKitScenes: A diverse real-world dataset for 3D indoor scene understanding using mobile RGB-D data. *NeurIPS Datasets and Benchmarks*, 2(6), 16.

Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A., 2023. Objaverse: A universe of annotated 3D objects. Proc. *CVPR*, 13142–13153.

Gao, R., Qi, Y., 2024. A Brief Review on Differentiable Rendering: Recent Advances and Challenges. Electronics, 13(17), 3546.

Greff, K., Belletti, F., Beyer, L., Doersch, C., Du, Y., Duckworth, D., Fleet, D. J., Gnanapragasam, D., Golemo, F., Herrmann, C. et al., 2022. Kubric: A scalable dataset generator. Proc. *CVPR*, 3749–3761.

Huang, P.-H., Matzen, K., Kopf, J., Ahuja, N., Huang, J.-B., 2018. Deepmvs: Learning multi-view stereopsis. Proc. *CVPR*, 2821–2830.

Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N., Luiten, J., Lopez-Antequera, M., Rota Bulò, S., Richardt, C., Ramanan, D., Scherer, S., Kontschieder, P., 2025. MapAnything: Universal feed-forward metric 3D reconstruction. *arXiv:2509.13414*.

Leroy, D., V., Cabon, Y., Revaud, J., 2024. Grounding image matching in 3d with MASt3R. Proc. *ECCV*.

Kendall, A., Grimes, M., Cipolla, R., 2015. PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization. Proc. *ICCV*.

Li, Z., Snavely, N., 2018. Megadepth: Learning single-view depth prediction from internet photos. Proc. *CVPR*, 2041- 2050.

Ling, L., Sheng, Y., Tu, Z., Zhao, W., Xin, C., Wan, K., Yu, L., Guo, Q., Yu, Z., Lu, Y. et al., 2024. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. Proc. *CVPR*, 22160-22169.

Morelli, L., Ioli, F., Maiwald, F., Mazzacca, G., Menna, F., Remondino, F., 2024. Deep-image-matching: a toolbox for multiview image matching of complex scenarios. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLVIII-2/W4-2024, 309-316.

Mu, T.J., Chen, H.X., Cai, J.X., Guo, N., 2023. Neural 3D reconstruction from sparse views using geometric priors. *Comp. Visual Media*, 9, 687-697.

Navaneet, K.L., Mandikal, P., Jampani, V., Babu, R.V., 2019. DIFFER: moving beyond 3d reconstruction with differentiable feature rendering. Proc. *CVPR*.

Nex, F., Stathopoulou, E. K., Remondino, F., Yang, M. Y., Madhuanand, L., Yogender, Y., et al., 2024. UseGeo-A UAV-based multi-sensor dataset for geospatial research. *ISPRS Open Journal of Photogrammetry and Remote Sensing, 13, 100070*.

Pan, X., Charron, N., Yang, Y., Peters, S., Whelan, T., Kong, C., Parkhi, O., Newcombe, R., Ren, Y. C., 2023. Aria digital twin: A new benchmark dataset for egocentric 3d machine percep- tion. Proc. *ICCV*, 20133-20143.

Perda, G., Morelli, L., Remondino, F., 2025. Orientation of ambiguous image sequences with similar and repeated structures. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci*., XLVIII-1/W4-2025, 95-102.

Pollefeys, M., Van Gool, L., Vergauwen, M., Verbiest, F., Cornelis, K., Tops, J., Koch, R., 2004. Visual modeling with a hand-held camera. *Int. Journal of Computer Vision*, 59, pp. 207-232.

Reizenstein, J., Shapovalov, R., Henzler, P., Sbordone, L., Labatut, P., Novotny, D., 2021. Common objects in 3D: Large-scale learning and evaluation of real-life 3D category reconstruction. Proc. *ICCV*, 10901-10911.

Remondino, F., El-Hakim, S., 2006. Image-based 3D modelling: a review. *The Photogrammetric Record*, Vol.21(115), September 2006, pp. 269-291.

Remondino, F., Nocerino, E., Toschi, I., Menna, F., 2017. A critical review of automated photogrammetric processing of large datasets. *ISPRS Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.,* Vol. XLII-2/W5, pp. 591-599.

Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M. A., Paczan, N., Webb, R., Susskind, J. M., 2021. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. Proc. *ICCV*, 10912-10922.

Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J. et al., 2019. Habitat: A platform for embodied ai research. Proc. *ICCV*, 9339-9347.

Schonberger, J.L. and Frahm, J.-M., 2016. Structure-from-Motion Revisited. Proc. *CVPR*.

Shi, Z., Meng, Z., Xing, Y., Ma, Y. and Wattenhofer, R., 2021. 3d-retr: End-to-end single and multi-view 3D reconstruction with transformers. *arXiv preprint arXiv:2110.08861.*

Stathopoulou, E.K., Remondino, F., 2023. A survey of conventional and learning-based methods for multi-view stereo. *The Photogrammetric Record*, Vol. 38(183), pp. 374-407.

Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J. J., Mur-Artal, R., Ren, C., Verma, S. et al., 2019. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv*:1906.05797.

Wang, H., Agapito, L., 2025. 3D reconstruction with spatial memory. Proc. *3DV*, 78-89.

Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D., 2025a. VGGT: Visual geometry grounded transformer. Proc. *CVPR*, 5294-5306.

Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T., 2025b. Permutation-Equivariant Visual Geometry Learning. *arXiv:2507.13347*.

Wang, J., Karaev, N., Rupprecht, C., Novotny, D., 2024a. VGGSfM: Visual geometry grounded deep structure from motion. Proc. *CVPR*, 21686-21697.

Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J., 2024b. DUSt3R: Geometric 3D vision made easy. Proc. *CVPR*, 20697–20709.

Wang, P., Tan, H., Bi, S., Xu, Y., Luan, F., Sunkavalli, K., Wang, W., Xu, Z., Zhang, K., 2024c. PF-LRM: Pose-free large reconstruction model for joint pose and shape prediction. Proc. *ICLR*, 2024.

Wang, X., Zhu, Z., Huang, G., Qin, F., Ye, Y., He, Y., Chi, X., Wang, X., 2022. MVSTER: Epipolar transformer for efficient multi-view stereo. Proc. *ECCV*, pp. 573-591.

Xia, H., Fu, Y., Liu, S., Wang, X., 2024. RGBD objects in the wild: Scaling real-world 3d object learning from RGB-D videos. Proc. *CVPR*, 22378-22389.

Yan, Z., Mazzacca, G., Rigon, S., Farella, E. M., Trybala, P., Remondino, F. et al., 2023. NeRFBK: a holistic dataset for benchmarking NeRF-based 3D reconstruction. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48(1), 219-226.

Yang, J., Sax, A., Liang, K. J., Henaff, M., Tang, H., Cao, A., Chai, J., Meier, F., Feiszli, M., 2025. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. Proc. *CVPR*, 21924-21935.

Yao, Y., Luo, Z., Li, S., Zhang, J., Ren, Y., Zhou, L., Fang, T., Quan, L., 2020a. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. Proc. *CVPR*, 1790-1799.

Yao, Y., Luo, Z., Li, S., Zhang, J., Ren, Y., Zhou, L., Fang, T., Quan, L., 2020b. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. Proc. *CVPR*, 1790-1799.

Yao, G., Yilmaz, A., Meng, F. and Zhang, L., 2021. Review of wide-baseline stereo image matching based on deep learning. *Remote Sensing*, 13(16), p.3247.

Yeshwanth, C., Liu, Y.-C., Niessner, M., Dai, A., 2023. Scannet++: A high-fidelity dataset of 3d indoor scenes. Proc. *ICCV*, 12–22.

Yin, R., Karaoglul, S., Gevers, T., 2023. Geometry-guided feature learning and fusion for indoor scene reconstruction. Proc. *ICCV*.

Zheng, Y., Harley, A. W., Shen, B., Wetzstein, G., Guibas, L. J., 2023. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. Proc. *ICCV*, 19855-19865.

Zhang, J., Li, Y., Chen, A., Xu, M., Liu, K., Wang, J., Long, X.-X., Liang, H., Xu, Z., Su, H. et al., 2025. Advances in feed-forward 3D reconstruction and view synthesis: A survey. *arXiv*:2507.14501.