AI-Based Camera Pose Estimation on Mixed Aerial and Ground Images: A Comparative Study

Zichao Zeng*, June Moh Goo, Jan Boehm

Department of Civil, Environmental and Geomatic Engineering, University College London, Gower Street, London, WC1E 6BT UK - (zichao.zeng.21, june.goo.21, j.boehm)@ucl.ac.uk

Keywords: Pose Estimation, Image Orientation, Cross-View, Foundation Model, Bundle Adjustment, Aerial-Ground Reconstruction

Abstract

Estimating camera poses jointly from aerial and ground imagery remains difficult because large viewpoint changes reduce overlap, alter appearance, and weaken the geometric assumptions relied on by both classical photogrammetry and recent AI-based reconstruction models. This paper presents a controlled comparison between a classic photogrammetric approach represented by COLMAP and a cross-view fine-tuned end-to-end model based on Dust3R. Tests are carried out on a London building scene containing 10 aerial and 29 ground images. Fine-tuned Dust3R reconstructs the full image set, whereas COLMAP successfully registers 24 ground-level images. Because both reconstructions are defined only up to an unknown similarity transform and no ground-truth poses are available, we evaluate the shared subset through 7-DoF similarity transformation analysis rather than direct metric pose errors. After transformation, the translation RMSE of the shared camera centres is 10.0% of the reconstructed scene diagonal in the fine-tuned Dust3R coordinate frame. We further compare pairwise geometric support using a unified fundamental-matrix RANSAC evaluation over 406 image pairs. The AI-based pipeline achieves substantially higher inlier ratios than photogrammetric pipeline under the same verification settings, indicating more successful cross-view orientation. The study contributes a clearer evaluation protocol for mixed aerial-ground pose estimation without ground truth, together with an empirical analysis of robustness, alignment behaviour, and current limitations of both pipelines.

1. Introduction

Aerial-ground reconstruction is increasingly relevant to urban modelling, mapping, robotics, and geospatial scene understanding (Brosh et al., 2019; Zheng et al., 2026; Berrabah et al., 2011; Hou et al., 2025; Zhang et al., 2026) because aerial and ground viewpoints provide complementary geometric coverage (Nex et al., 2015). A prerequisite for such fusion is reliable camera pose estimation across images whose viewing directions, scales, and visible content may differ substantially. In practice, this is difficult: aerial images often observe roofs and upper facades, while ground images mainly capture street-level facades and occluded scene regions. The resulting reduction in overlap makes mixed aerial-ground pose estimation much less stable than conventional same-modality reconstruction.

Classical photogrammetric pipelines (Luhmann et al., 2023; Schonberger and Frahm, 2016; Schönberger et al., 2016) remain the standard solution for image orientation and 3D reconstruction. Their feature matching, geometric verification, and bundle adjustment framework can deliver high-quality poses when image overlap is sufficient. However, under strong aerial-ground viewpoint changes, the number of geometrically verifiable matches may drop sharply, which can prevent image registration altogether. Recent AI-based reconstruction models such as Dust3R (Wang et al., 2024b), MASt3R (Leroy et al., 2024), and related methods (Yang et al., 2025; Wang et al., 2025) offer a different operating regime by predicting dense cross-view geometry from learned visual priors, and fine-tuning on cross-view data (Vuong et al., 2025) has started to make these models viable for aerial-ground settings.

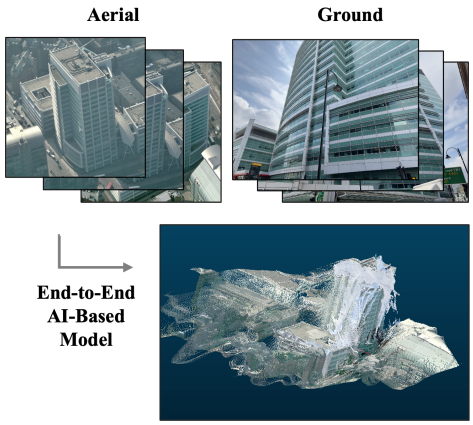
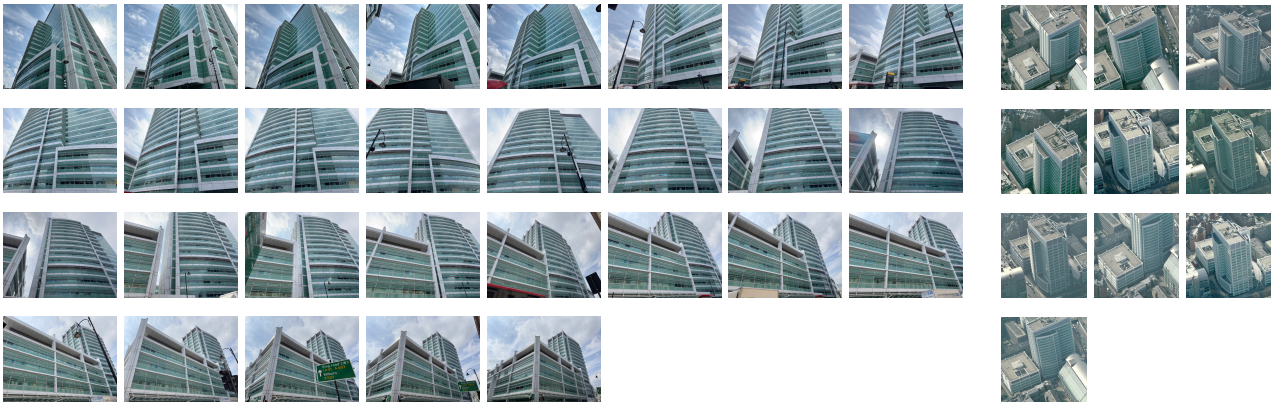


Figure 1. Aerial+Ground 3D Reconstruction

This creates a practical question: when classical photogrammetric and AI-based reconstruction behave very differently on the same mixed-view dataset, how should they be compared fairly? Counting only the number of registered images is insufficient, because a method may recover more poses while still producing a trajectory that is difficult to interpret geometrically. Conversely, a more conservative pipeline may register fewer images but maintain strong local consistency where it succeeds. The problem is further complicated by the absence of ground-truth camera poses in many real scenes and by the fact that both reconstructions are often defined only up to an unknown similarity transform.

In this paper, we compare default COLMAP with a cross-view fine-tuned Dust3R model on a real mixed aerial-ground im-

* Corresponding author



(a) Ground views (b) Aerial views

Figure 2. Input dataset consisting of terrestrial image data and oblique airborne data.

age set. Since no external metric reference is available, we focus on relative evaluation. We first align the shared subset of recovered camera centres through a 7-DoF similarity transformation (Umeyama, 2002) and analyse translation agreement in both absolute model units and a scale-independent normalised form. We then evaluate pairwise geometric support using a common fundamental-matrix RANSAC backend applied to correspondences from both pipelines. In addition, we inspect pose visualisations for the shared ground-level subset and for the full reconstructions produced by each method. Our contributions are as follows:

- A cleaner relative evaluation protocol for comparing photogrammetric and AI-based camera pose estimation on mixed aerial-ground imagery without ground-truth camera poses.
- A quantitative analysis of shared-pose agreement based on 7-DoF alignment, translation accuracy, and scale-independent pose error.
- An empirical study showing how fine-tuned Dust3R and COLMAP differ in registration coverage, pairwise inlier support, and pose behaviour on a challenging cross-view scene.

2. Related Work

Aerial–ground pose estimation Estimating camera poses from mixed aerial and ground imagery is challenging due to large viewpoint differences, occlusions, and limited co-visible regions between nadir or oblique aerial views and ground-level imagery (Nex et al., 2015; Zheng et al., 2026). At the same time, aerial–ground fusion is increasingly important for urban modelling, mapping, robotics, and geospatial perception (Brosh et al., 2019; Hou et al., 2025; Zhang et al., 2026; Berrabah et al., 2011), since combining complementary viewpoints enables more complete scene understanding than either modality alone (Fruh and Zakhor, 2003; Zheng et al., 2020). Prior research (Hu et al., 2018; Durgam et al., 2024) has explored cross-view localisation, cross-view matching, and aerial–ground registration, but most works target pairwise localisation or cross-view retrieval rather than full multi-view pose estimation. Datasets specifically designed for evaluating pose estimation under aerial–ground conditions remain scarce.

Photogrammetric multi-view reconstruction Traditional photogrammetry remains a mature and reliable methodology for 3D reconstruction (Schonberger and Frahm, 2016; Luhmann et al., 2023; Vijayan and Pushpalatha, 2020; Schönberger et al., 2016), employing classical features (e.g., SIFT), robust matching, geometric filtering, and bundle adjustment (BA). These methods deliver high-accuracy camera poses in aerial mapping and ground surveys when image overlap and viewpoint consistency are sufficient. Their robustness, however, decreases when viewpoint disparities increase, as the number of geometrically verifiable matches between aerial and ground viewpoints becomes limited (Schonberger et al., 2017; Mikolajczyk and Schmid, 2005). Although learned features can improve matching resilience (DeTone et al., 2018; Sarlin et al., 2020), there has been little systematic evaluation of how classical photogrammetry compares against modern learned or AI-based approaches specifically for aerial–ground pose estimation. Most prior comparisons consider generic multi-view datasets or indoor/outdoor scenes (Huang et al., n.d.), rather than structured cross-view conditions.

AI-based multi-view reconstruction Recent end-to-end models such as DUSt3R (Wang et al., 2024b), MASt3R (Leroy et al., 2024), VGGSfM (Wang et al., 2024a), and Fast3R (Yang et al., 2025) predict camera geometry directly from dense learned correspondences or global image representations (Goo et al., 2025). These models show strong performance and robustness on general multi-view benchmarks and wide-baseline datasets. However, their generalisation to aerial–ground imagery is limited without domain-specific fine-tuning (Durgam et al., 2024). Comparative understanding remains limited: few works evaluate AI-based pose estimators against classical photogrammetry on identical aerial–ground image sets, and even fewer examine their geometric consistency in the absence of ground-truth poses. This gap motivates the relative, similarity-based comparison presented in this study.

3. Dataset and Experimental Setup

We use a London building scene as a case study. The dataset contains 39 images in total: 10 oblique aerial views and 29 ground-level views. Figure 2 shows the input images, and Figure 1 illustrates the mixed aerial-ground reconstruction setting. The scene is intentionally challenging because the aerial and street-level views share only partial visible content and exhibit strong perspective changes.

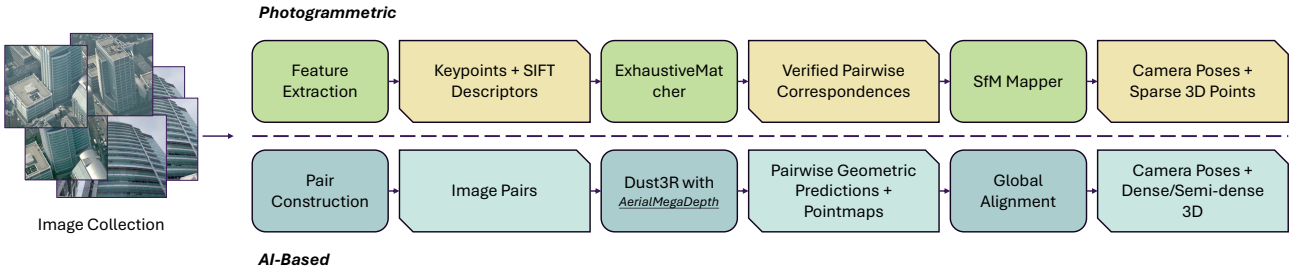


Figure 3. Overview of the comparison pipeline between the AI-based model (Dust3R fine-tuned on AerialMegaDepth) and the classical photogrammetric pipeline (default COLMAP).

Parameter	Value
Scale s	0.028
Alignment rotation $\mathbf{R}_{\text{align}}$	$\begin{bmatrix} 0.791 & -0.589 & -0.164 \\ -0.432 & -0.347 & -0.832 \\ 0.434 & 0.729 & -0.529 \end{bmatrix}$
Alignment translation $\mathbf{t}_{\text{align}}$	$\begin{bmatrix} 0.067 \\ 0.102 \\ 1.378 \end{bmatrix}$

Table 1. Estimated 7-DoF similarity transformation parameters $(s, \mathbf{R}_{\text{align}}, \mathbf{t}_{\text{align}})$ mapping COLMAP camera centres into the Dust3R frame.

Scene diagonal length	0.388
Relative translation RMSE	0.100
Approx. relative accuracy	1:10

Table 2. Scale-independent translation relative accuracy obtained by normalising the RMSE with the reconstructed scene diagonal.

Two reconstruction pipelines are evaluated on this dataset. The first is default COLMAP (Schonberger and Frahm, 2016; Schönberger et al., 2016), representing a standard photogrammetric SfM workflow based on sparse feature matching, geometric verification, and bundle adjustment. The second is Dust3R (Wang et al., 2024b) fine-tuned on AerialMegaDepth (Vuong et al., 2025), representing an AI-based model adapted specifically for cross-view aerial-ground geometry. Dust3R is applied to the full 39-image set and produces poses for all images. COLMAP successfully registers 24 images, all from the ground-level subset, while the remaining images are not oriented within the same reconstruction.

Because the two methods do not reconstruct the same set of cameras, the quantitative comparison must be restricted to the shared subset. We therefore evaluate pose agreement on the 24 ground-level images that are oriented by both pipelines. This restriction is important for interpretation: the shared-subset analysis measures agreement where COLMAP succeeds, but it does not capture the difficulty of the images rejected by COLMAP. We therefore complement the numerical analysis with all pose visualisations for each method.

3.1 Evaluation Protocol

We use three complementary views of performance: shared-trajectory alignment, controlled pairwise geometric verification, and qualitative trajectory visualisation. Together, these analyses capture global agreement on the shared subset, local geometric verifiability of correspondences, and reconstruction coverage beyond the shared subset, all without requiring ground-truth camera poses.

Method	Median	Mean
COLMAP	0.160	0.167
Fine-tuned Dust3R	0.502	0.488

Table 3. RANSAC inlier ratios across 406 image pairs under a shared fundamental-matrix verification pipeline.

First, the camera centres from the shared subset are aligned using a 7-DoF similarity transformation. Let $\mathbf{c}_i^A \in \mathbb{R}^3$ and $\mathbf{c}_i^B \in \mathbb{R}^3$ denote the camera centres of two methods for the same image i. The similarity transformation is estimated by minimising the following objective:

$$\min_{s,\mathbf{R}_{\text{align}},\mathbf{t}_{\text{align}}} \sum_{i=1}^{N} \left\| \mathbf{c}_i^A - \left(s\mathbf{R}_{\text{align}}\mathbf{c}_i^B + \mathbf{t}_{\text{align}} \right) \right\|_2^2, \quad (1)$$

where $s \in \mathbb{R}^+$ is the global scale, $\mathbf{R}_{\text{align}} \in SO(3)$ is the alignment rotation, and $\mathbf{t}_{\text{align}} \in \mathbb{R}^3$ is the alignment translation. The aligned camera centre of method B is

$$\hat{\mathbf{c}}_i^B = s\mathbf{R}_{\text{align}}\mathbf{c}_i^B + \mathbf{t}_{\text{align}}, \quad (2)$$

and the per-image translation error is computed as

$$e_i^{\text{trans}} = \left\| \mathbf{c}_i^A - \hat{\mathbf{c}}_i^B \right\|_2, \quad (3)$$

while the translation RMSE over the shared subset is

$$e_{\text{RMSE}}^{\text{trans}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(e_i^{\text{trans}})^2}. \quad (4)$$

To make this quantity scale-independent, we normalise it by the scene diagonal d_{scene}:

$$e_{\text{rel}}^{\text{trans}} = \frac{e_{\text{RMSE}}^{\text{trans}}}{d_{\text{scene}}}. \quad (5)$$

We report the estimated alignment parameters $(s, \mathbf{R}_{\text{align}}, \mathbf{t}_{\text{align}})$, together with the translation statistics derived from $\{e_i^{\text{trans}}\}_{i=1}^N$: the RMSE $e_{\text{RMSE}}^{\text{trans}}$, the median error, the maximum error, and the relative translation error $e_{\text{rel}}^{\text{trans}}$. Since both reconstructions have unknown scale and no shared metric reference, the similarity transform is used

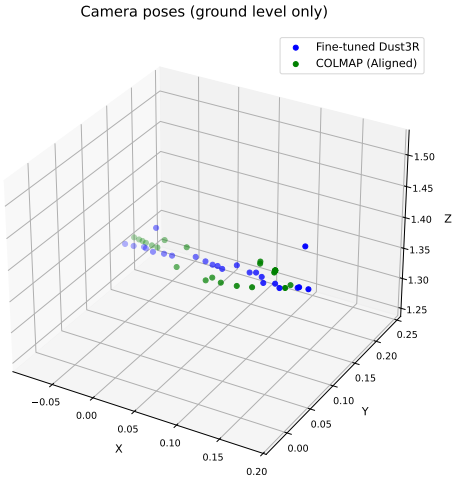


Figure 4. Aligned camera poses for the 24 shared ground-level cameras reconstructed by both Dust3R and COLMAP.

only to register the trajectories into a common frame. The alignment rotation $\mathbf{R}_{\text{align}}$ is therefore reported as part of the transformation itself, but not interpreted as a standalone pose-accuracy metric.

Second, we perform a controlled RANSAC inlier-ratio comparison on pairwise correspondences. For each image pair, the inlier ratio is defined as

$$r_{\text{inlier}} = \frac{N_{\text{inlier}}}{N_{\text{match}}}, \tag{6}$$

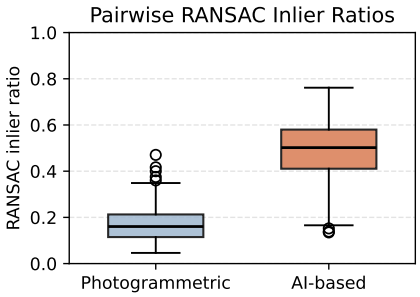
where N_{match} is the total number of tentative correspondences and N_{inlier} is the number classified as inliers by a common fundamental-matrix RANSAC solver. Both the AI-based and COLMAP-derived correspondences are evaluated with the same OpenCV implementation, reprojection threshold, and RANSAC hyperparameters. This ensures that the comparison reflects the geometric verifiability of the correspondences themselves rather than differences in the verification backend.

Third, we analyse trajectory visualisations to characterise reconstruction coverage and overall spatial behaviour. This qualitative comparison is performed both on the shared ground-level subset after similarity transformation and on the complete trajectories reconstructed independently by each method. Its purpose is not to define an additional scalar error metric, but to reveal whether the aligned trajectories follow comparable spatial trends on the shared subset and to show how the two pipelines differ in the extent of the scene they are able to reconstruct.

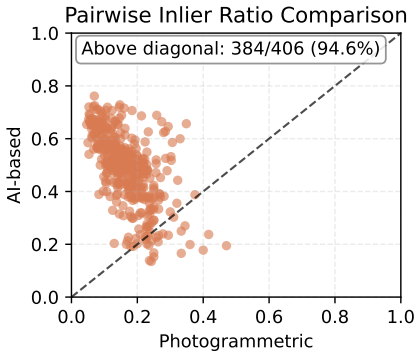
4. Results and Discussion

Global pose alignment. Table 1 reports the estimated alignment parameters $(s, \mathbf{R}_{\text{align}}, \mathbf{t}_{\text{align}})$ that map the COLMAP camera centres into the Dust3R coordinate frame. These quantities should be understood only as the similarity transformation needed to reconcile two independent reconstructions with different gauge freedoms. Because both reconstructions are defined up to an unknown scale and do not share a common metric reference, $\mathbf{R}_{\text{align}}$ is not interpreted as a direct pose-accuracy indicator.

The shared-subset agreement is therefore evaluated through the translation statistics defined in the protocol. Table 2 reports



(a) Boxplot of per-pair inlier ratios.



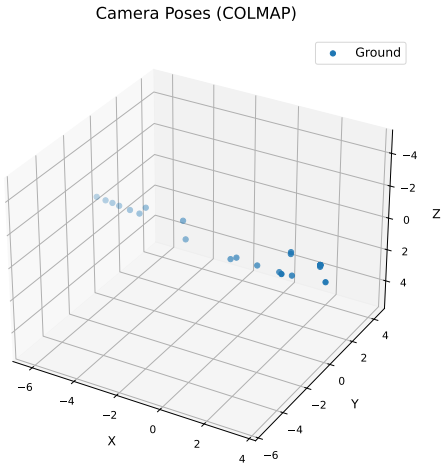
(b) Pairwise comparison of COLMAP and fine-tuned Dust3R inlier ratios.

Figure 5. RANSAC inlier-ratio comparison across the 406 evaluated image pairs.

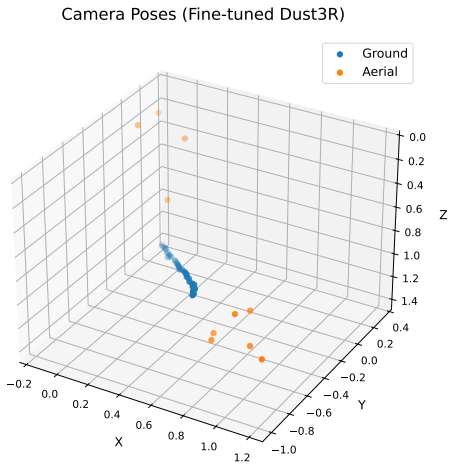
the corresponding relative translation error $e_{\text{rel}}^{\text{trans}} = 0.100$, obtained by normalising the RMSE with the scene diagonal. This scale-independent result suggests that the two pipelines recover a similar global spatial layout where both succeed, although the remaining discrepancy is still non-negligible with respect to the overall reconstructed extent.

Controlled RANSAC inlier-ratio comparison. Under the same verification settings, the AI-based correspondences produce clearly higher inlier support than COLMAP. Table 3 shows that Dust3R achieves a median inlier ratio of 0.502 and a mean of 0.488, compared with 0.160 and 0.167 for COLMAP. Figure 5 further illustrates the difference: the boxplot shows a much higher overall distribution for Dust3R, while the scatter plot indicates that most evaluated image pairs lie above the equality line, meaning that the AI-based correspondences are more verifiable under the same fundamental-matrix model on most pairs. This is a strong indication that the learned cross-view representation is better able to sustain tentative matching across large viewpoint changes.

Pose visualisations. The figures 4 complements the quantitative metrics by making the shared-subset agreement and whole-scene coverage visually interpretable. Figure 4 shows the aligned poses for the 24 shared ground-level cameras. The two paths follow a similar spatial trend after applying $(s, \mathbf{R}_{\text{align}}, \mathbf{t}_{\text{align}})$, which is consistent with the relatively small values of $e_{\text{RMSE}}^{\text{trans}}$ and $e_{\text{rel}}^{\text{trans}}$. Figure 6 extends this view to the complete reconstructions and reveals an important asymmetry between the methods: COLMAP only reconstructs ground-level poses, whereas Dust3R produces a single reconstruction spanning both ground and aerial viewpoints. This qualitative coverage analysis reinforces the practical interpretation of the



(a) COLMAP poses.



(b) Fine-tuned Dust3R poses.

Figure 6. Full reconstructed camera poses produced independently by COLMAP and fine-tuned Dust3R. COLMAP reconstructs only the ground-level subset, whereas fine-tuned Dust3R reconstructs both aerial and ground views.

numerical results, namely that Dust3R is substantially more robust on this mixed-view scene.

Overall interpretation. Taken together, the three analyses in the evaluation protocol suggest a complementary picture rather than a simple winner-takes-all conclusion. The alignment shows that, on the 24 common cameras, the two pipelines recover a broadly compatible spatial configuration after similarity transformation. The controlled inlier-ratio comparison shows that Dust3R provides much stronger pairwise geometric support under the same verifier. The trajectory visualisations then show that this stronger pairwise support translates into substantially broader reconstruction coverage, especially across the aerial-ground gap. The main conclusion is therefore that AI-based cross-view reconstruction offers a substantial advantage in coverage and match robustness for mixed aerial-ground imagery, while the shared-subset alignment remains a relative consistency check rather than a full measure of whole-scene geometric accuracy.

In addition, Figure 7 provides a qualitative comparison between the ground-truth camera poses and the aligned fine-tuned

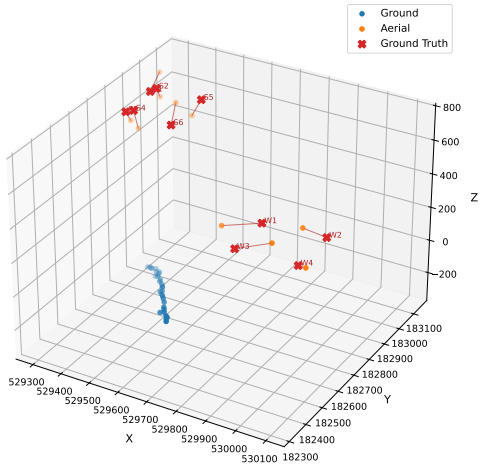


Figure 7. Comparison between ground-truth camera poses and aligned fine-tuned Dust3R camera poses for the full image set. Red line segments indicate the residual offset between each prediction and its reference pose.

Dust3R estimates over the full image set. The predicted poses generally follow the same global spatial configuration as the reference metadata, indicating that the model recovers a coherent camera layout across both aerial and ground views. The red connecting segments visualize the residual offsets between prediction and reference for each camera. While most cameras remain reasonably close after alignment, several larger offsets are still visible, suggesting local pose deviations in more challenging viewpoints. Therefore, the figure supports the interpretation that Dust3R captures the overall scene structure well, but residual camera-wise errors remain.

5. Conclusion and Limitations

This paper presented a comparison between a classical photogrammetric pipeline and a cross-view fine-tuned AI-based model for mixed aerial-ground pose estimation. On the London case study, Dust3R reconstructed all 39 images, whereas COLMAP registered only 24 ground-level views. After 7-DoF alignment on the shared subset, the two methods showed moderate translation agreement, and the AI-based correspondences achieved substantially higher RANSAC inlier ratios under a common verification setup. These results indicate that cross-view fine-tuned AI-based reconstruction is currently more robust than default photogrammetry for this challenging mixed-view scene.

The study also has clear limitations. First, there is no full ground-truth camera reference, so the evaluation remains relative and cannot provide metric pose accuracy. Second, the shared-subset trajectory analysis excludes the images that COLMAP fails to register, which introduces an unavoidable selection bias. Third, both reconstructions are only defined up to similarity, making direct interpretation of rotational agreement weak in the absence of a common metric frame. Finally, the analysis is based on a single scene, so the conclusions should not yet be generalised too broadly.

Future work will extend the study to additional scenes with different urban layouts and viewpoint gaps (Nex et al., 2015), include more AI-based and hybrid reconstruction

pipelines (Leroy et al., 2024; Yang et al., 2025; Wang et al., 2025), and incorporate stronger external references where available. A broader benchmark with scene-level statistics and failure-case analysis would make it possible to determine not only which method reconstructs more images, but also under what conditions that additional coverage translates into reliable geometry.

6. Acknowledgement

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

References

Berrabah, S. A., Sahli, H., Baudoin, Y., 2011. Visual-based simultaneous localization and mapping and global positioning system correction for geo-localization of a mobile robot. *Measurement Science and Technology*, 22(12), 124003.

Brosh, E., Friedmann, M., Kadar, I., Yitzhak Lavy, L., Levi, E., Rippa, S., Lempert, Y., Fernandez-Ruiz, B., Herzig, R., Darrell, T., 2019. Accurate visual localization for automotive applications. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 0–0.

DeTone, D., Malisiewicz, T., Rabinovich, A., 2018. Superpoint: Self-supervised interest point detection and description. *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 224–236.

Durgam, A., Paheding, S., Dhiman, V., Devabhaktuni, V., 2024. Cross-view geo-localization: a survey. *IEEE Access*.

Fruh, C., Zakhor, A., 2003. Constructing 3D city models by merging aerial and ground views. *IEEE computer graphics and applications*, 23(6), 52–61.

Goo, J. M., Zeng, Z., Morelli, L., Remondino, F., Boehm, J., 2025. Exploring modern end-to-end AI-based multi-view 3D reconstruction. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48, 91–97.

Hou, Q., Hou, C., Zhang, F., Weng, Q., 2025. Crowdsourced images geo-localization method based on multi-modal deep learning. *EGU General Assembly Conference Abstracts*, EGU25–14877.

Hu, S., Feng, M., Nguyen, R. M., Lee, G. H., 2018. Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7258–7267.

Huang, K. W., Li, B., Hariharan, B., Snavely, N., n.d. C3po: Cross-view cross-modality correspondence by pointmap prediction. *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Leroy, V., Cabon, Y., Revaud, J., 2024. Grounding image matching in 3d with mast3r.

Luhmann, T., Robson, S., Kyle, S., Boehm, J., 2023. *Close-range photogrammetry and 3D imaging*. Walter de Gruyter GmbH & Co KG.

Mikolajczyk, K., Schmid, C., 2005. A performance evaluation of local descriptors. *IEEE transactions on pattern analysis and machine intelligence*, 27(10), 1615–1630.

Nex, F., Gerke, M., Remondino, F., Przybilla, H.-J., Bäumker, M., Zurhorst, A., 2015. ISPRS benchmark for multi-platform photogrammetry. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2, 135–142.

Sarlin, P.-E., DeTone, D., Malisiewicz, T., Rabinovich, A., 2020. Superglue: Learning feature matching with graph neural networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Schonberger, J. L., Frahm, J.-M., 2016. Structure-from-motion revisited. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4104–4113.

Schonberger, J. L., Hardmeier, H., Sattler, T., Pollefeys, M., 2017. Comparative evaluation of hand-crafted and learned local features. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1482–1491.

Schönberger, J. L., Zheng, E., Pollefeys, M., Frahm, J.-M., 2016. Pixelwise view selection for unstructured multi-view stereo. *European Conference on Computer Vision (ECCV)*.

Umeyama, S., 2002. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 13(4), 376–380.

Vijayan, V., Pushpalatha, K., 2020. A comparative analysis of RootSIFT and SIFT methods for drowsy features extraction. *Procedia computer science*, 171, 436–445.

Vuong, K., Ghosh, A., Ramanan, D., Narasimhan, S., Tulsiani, S., 2025. Aerialmegadepth: Learning aerial-ground reconstruction and view synthesis. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 21674–21684.

Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D., 2025. Vggt: Visual geometry grounded transformer. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 5294–5306.

Wang, J., Karaev, N., Rupprecht, C., Novotny, D., 2024a. Vggsfm: Visual geometry grounded deep structure from motion. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 21686–21697.

Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J., 2024b. Dust3r: Geometric 3d vision made easy. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20697–20709.

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. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 21924–21935.

Zhang, Y., Ke, E., Kwan, M.-P., Fang, L., Li, M., 2026. Multi-frequency street-level urban noise modeling and mapping through street view and remote sensing image fusion. *Computers, Environment and Urban Systems*, 126, 102401.

Zheng, J., Dai, R., Liu, R., Zeng, Z., Chen, Y., Wang, F., Peng, K., Yang, K., Zhang, J., Stiefelhagen, R., 2026. RHO: Robust Holistic OSM-Based Metric Cross-View Geo-Localization. *arXiv preprint arXiv:2603.27758*.

Zheng, Z., Wei, Y., Yang, Y., 2020. University-1652: A multi-view multi-source benchmark for drone-based geo-localization. *Proceedings of the 28th ACM international conference on Multimedia*, 1395–1403.