Evaluation of VGGT with ALS Point Clouds for Large-Scale Dense Mapping

Yandi Yang [1], Naser El-Sheimy [1]

[1] Dept. of Geomatic Engineering, University of Calgary, Calgary, Alberta, Canada - (yandi.yang, elsheimy)@ucalgary.ca

Keywords: Cross-Modal, Airborne Laser Scanning, Mobile Mapping.

ABSTRACT:

We present a framework that integrates ground-level imagery with Airborne Laser Scanning (ALS) point clouds. While the Visual Geometry Grounded Transformer (VGGT) enables dense geometry estimation from uncalibrated images, its application is limited by non-metric results and high GPU requirements. By leveraging publicly available, georeferenced ALS point clouds as an external metric constraint, our system restores absolute scale and global coordinates without requiring high-grade GNSS/INS or expensive on-board LiDAR systems. We introduce a confidence-weighted Sim (3) registration algorithm that utilizes a learned confidence mask to filter out unreliable points in dense street-level reconstructions. Experimental evaluations conducted on large-scale urban datasets demonstrate the average check point errors of 0.77 meters in Hong Kong dataset and 0.69 meters in Wuhan dataset, showing great potentials of feed-forward models in large-scale outdoor dense mapping.

1. INTRODUCTION

Accurate large-scale dense 3D reconstruction is fundamental for mobile mapping (Schwarz et al., 2004), robotics (Li et al., 2024, Li et al., 2026) and autonomous navigation (El-Sheimy et al., 2020). Traditionally, the generation of high-fidelity 3D models relies on mobile mapping systems (MMS) equipped with high-grade LiDAR and GNSS/INS sensors. While effective, these systems are expensive and intensive for frequent updates. Structure from Motion (Wang et al., 2024) and visual SLAM (Murai et al., 2025) workflows offer a more accessible method by reconstructing environments from ground-level imagery; however, they suffer from cumulative scale drift over kilometer-scale trajectories and typically require precise camera calibration or expensive ground control points (GCPs) to maintain global consistency.

Recent advances in feed-forward 3D reconstruction, notably the Visual Geometry Grounded Transformer (VGGT) (Wang et al., 2025), have demonstrated the ability to generate dense point clouds directly from RGB images without explicit optimization. VGGT jointly estimates camera poses, depth, and dense geometry from multiple uncalibrated frames in a single forward pass. However, its scalability is limited by two factors: (1) the lack of absolute metric scale as the output point clouds are in a normalized space and (2) high GPU memory demands which leads to alignment errors when stitching long sequences. Several works have been proposed to extend VGGT to large-scale applications. VGGT-Long (Deng et al., 2025) and VGGT-SLAM (Maggio et al., 2025, Maggio et al., 2026) partition sequences into submaps using offline and online schemes, respectively. However, they do not recover the metric scale of the point clouds. LiDAR-VGGT (Wang et al., 2026) and VGGT-MPR (Xu et al., 2026) fuse ground LiDAR with camera using VGGT which rely on on-board LiDAR-visual systems and suffer from global trajectory drift.

Many Image-to-Point Cloud algorithms have been proposed to address the cross-modal localization problem. CoFiI2P (Kang et al., 2024) proposes a coarse-to-fine pipeline for image to point cloud registration. Line features from images and point clouds are extracted for a Visual SLAM system in a prior point cloud map (Zheng et al., 2024). HyperMap (Chang et al., 2021) performs 3D sparse convolution for extracting and compressing features from point cloud maps. An air-ground image to point cloud dataset is proposed and benchmarked (Yang et al., 2026).

Many national mapping agencies (e.g., USGS, IGN, Ordnance Survey) have released Airborne Laser Scanning (ALS) datasets covering vast urban and rural areas. These high-quality aerial point clouds provide globally consistent, metrically referenced data that can serve as external constraints for ground-level reconstructions. However, they often lack the street-level density and colors.

In this work, we propose a novel framework that bridges the gap between ground-level feed-forward reconstruction results and aerial metric prior point clouds. Instead of relying on ground-based LiDAR or high-precision navigation system, we leverage ubiquitous ALS point clouds as external constraints to rectify and scale the local reconstructions from VGGT. The challenge lies in the significant modality and viewpoint gap: ALS provides sparse, top-down structures, while VGGT produces dense, street-level information. To address this, we introduce a coarse-to-fine registration pipeline that aligns these multi-source point clouds.

The main contributions of this article are as follows:

(1) A Cross-Modal Fusion Framework: We present a system that tightly integrates ground-level imagery with publicly available ALS data. This approach restores the absolute metric scale and global coordinates of VGGT reconstructions, eliminating the dependency on LiDAR or high-end GNSS/INS units.

(2) Confidence-Weighted Sim (3) Registration: To handle the differences in viewpoint, density, and coverage between aerial and terrestrial data, we propose a robust registration algorithm by incorporating a confidence mask.

(3) Large-Scale Datasets Validation: We validate our framework on dense urban datasets. Experimental results demonstrate great potential of VGGT in the application of large-scale outdoor dense mapping.

2. METHODOLOGY

In order to use ALS point clouds for global optimization of VGGT results, we partition the raw images into different sequences and process them with different submaps. Then for

each submap, we initialize the scale first by using the façade features between cross-source point clouds. After the initial alignment for each submap, we combine and optimize all the submaps with ALS point clouds.

2.1 Submap-based Processing

To handle long image sequences and memory limitations, the input ground-level imagery set $\mathbb{I}_N = \{I_1, I_2, \dots, I_N\}$ is partitioned into K overlapping submaps $\mathbb{S}_K = \{S_1, S_2, \dots, S_K\}$. Each submap is independently processed by the Visual Geometry Grounded Transformer (VGGT) to produce a dense point cloud P_k, camera poses T_k, and confidence maps C_k:

$$f(S_k) = \{P_k, T_k, C_k\}, k = 1,2, \dots, K. \quad (1)$$

Here, $P_k \in \mathbb{R}^{H\times W\times 3}$ represents the dense reconstruction for submap S_k, and $C_k \in \mathbb{R}^{H\times W}$ encodes the model's confidence for each point, T_k is the output extrinsic. Since each submap is reconstructed in its own local coordinate system, the resulting point clouds and extrinsic are not metrically scaled or globally aligned. The target is to recover the point clouds and extrinsic into real scales using prior ALS point clouds.
Each T_k is a Sim (3) transformation which contains a scale s, a rotation matrix $\boldsymbol{R}$ and a translation vector $\boldsymbol{t}$.

$$T_k = \begin{bmatrix} s\boldsymbol{R} & \boldsymbol{t} \\ 0 & 1 \end{bmatrix} \in Sim(3) \quad (2)$$

The optimization parameters are the Sim (3) transformation matrices for all the submaps by matching VGGT point clouds with ALS point clouds.

2.2 Alignment with ALS

To restore absolute scale and global reference, each VGGT submap is co-registered with the corresponding region of the Airborne Laser Scanning (ALS) data. Considering raw image point clouds from VGGT results have noises, and the most salient point cloud features in large-scale outdoor environments are building boundaries and façades, we first extract façade features from the VGGT point clouds and establish correspondences within overlapping areas to recover the initial scale and transformation matrix. To acquire reliable building features, we calculate the verticality of each point in the image point clouds by calculating covariance matrix with eigenvalues of $\lambda_1 > \lambda_2 > \lambda_3$ and eigenvectors v_1, v_2, v_3.The verticality V_λ is calculated as:

$$V_\lambda = 1 - \|z_{v_3}\| \quad (3)$$

z_{v_3} is the z value of v_3. Red points in Fig. 1 are the façade points with very high verticality, which are then used for initial alignment between image point clouds and ALS point clouds. An initial similarity transformation $T_k^{\text{init}} \in \text{Sim}(3)$ is then estimated using a confidence-weighted least squares formulation:

$$\mathbf{T}_k^{init} = \arg\min_{\mathbf{T}\in \text{Sim}(3)} \sum_{i\in\mathcal{M}_k} \mathbf{w}_i \left\| \mathbf{p}_i^{ALS} - \mathbf{T}\cdot \mathbf{p}_i^{VGGT} \right\|_2^2 \quad (4)$$

where M_k is the set of matched correspondences and the weight w_i incorporates the VGGT confidence value c_i, thereby down-weighting unreliable points (e.g., vehicles or pedestrians) while emphasizing stable structures such as façades and road surfaces.

The transformation includes rotation $R \in SO(3)$, translation $t \in \mathbb{R}^3$, and uniform scale $s \in \mathbb{R}^+$. Each alignment is efficiently solved in closed form using a weighted Umeyama algorithm. Fig. 1 shows the raw image with the initial reconstructed point clouds in real scale and color.

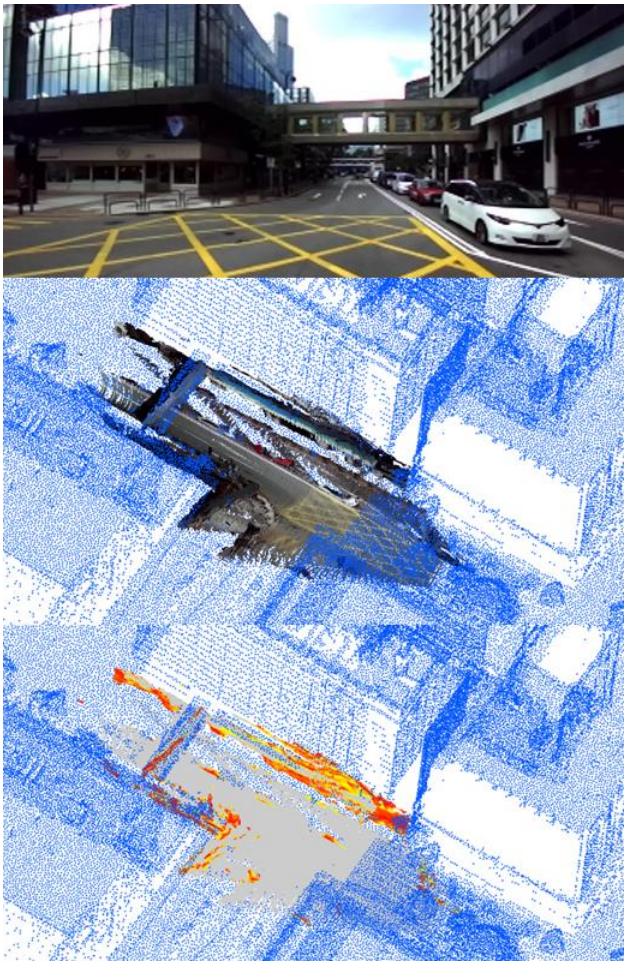

Figure 1. Façade features of VGGT point clouds.

2.3 Inter-Submap Alignment and Global Optimization

After the initial alignment with ALS point clouds, we have the initial scales and poses for each individual submap. Considering overlapping regions between adjacent submaps provide additional spatial constraints, we form a global factor graph to further improve the image point cloud quality by minimizing the submap-to-submap and submap-to-ALS point errors.

A relative transformation $T_{k,k+1}$ is estimated by minimizing the point-to-point distance between overlapping areas of neighboring submaps:

$$\mathbf{T}_{k,k+1} = \arg\min_{\mathbf{T}\in \text{Sim}(3)} \sum_{i\in O_{k,k+1}} \left\| \mathbf{p}_i^{k+1} - \mathbf{T}\cdot \mathbf{p}_i^{k} \right\|^2 \quad (5)$$

where $O_{k,k+1}$ denotes the set of overlapping points.

ALS point clouds can serve as prior information to eliminate global drifts of image point clouds. In order to achieve global point cloud consistency, all Sim(3) transformations are refined through pose-graph optimization that jointly enforces sequential (inter-submap) and submap-to-ALS alignment constraints:

https://doi.org/10.5194/isprs-archives-XLIX-B1-2026-77-2026

$$\left\{\mathbf{T}_k^{\text{global}}\right\}_{k=1}^{K} = \arg\min_{\{\mathbf{T}_k\}} \sum_{k=1}^{K-1} \left\|\log_{\text{Sim}(3)}\left(\mathbf{T}_{k,k+1}^{-1}\mathbf{T}_k^{-1}\mathbf{T}_{k+1}\right)\right\|^2 + \sum_{k=1}^{K} \left\|\log_{\text{Sim}(3)}\left((\mathbf{T}_k^{ALS})^{-1}\mathbf{T}_k\right)\right\|^2 \quad (6)$$

where $\log_{\text{Sim}(3)}(\cdot)$ maps each Sim(3) transformation into its tangent space for optimization on the Lie algebra and $\mathbf{T}_k^{ALS}$ is the registration results between each submap and ALS point clouds. The resulting transformations provide a globally consistent, metrically scaled dense reconstruction aligned to the ALS reference frame. Fig. 2 shows the process of inter-submap matching and global matching with ALS point clouds. Red points are image point clouds from a neighboring submap.

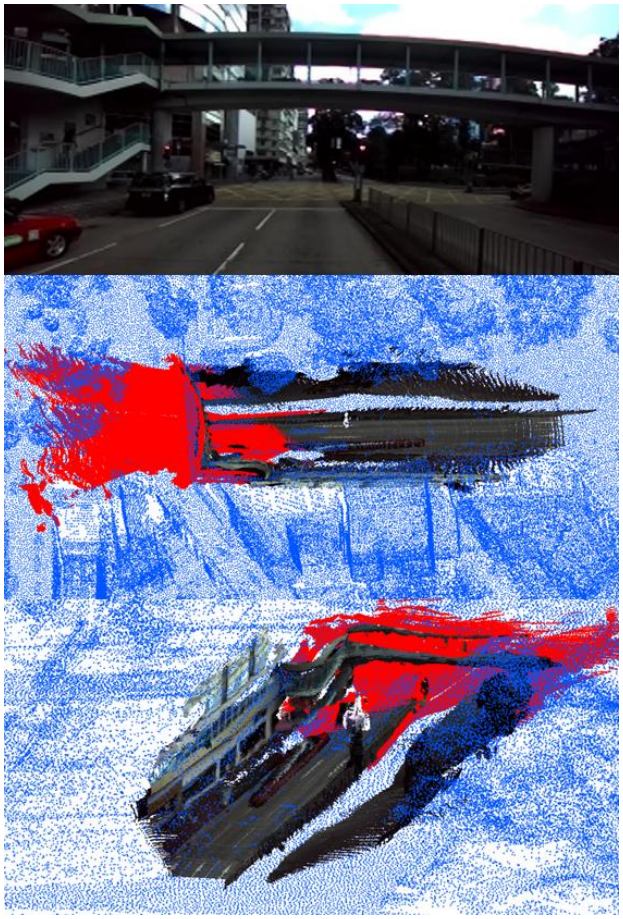

Figure 2. Inter-submap alignment and global optimization with ALS point clouds. Red points are overlapping points from a neighboring submap.

3. EXPERIMENTS

In order to test the applicability of the algorithm, we test the algorithm on two large-scale outdoor sequences (Hong Kong medium and Wuhan Loop 2) of AGI2P dataset (Yang et al., 2026). Then we evaluate the image point cloud accuracy by evenly selecting check points on ALS point clouds and calculating the corresponding check point offsets between image point clouds and ALS point clouds. We introduce the study area firstly and then evaluate the point cloud accuracy. We use NVIDIA-H100 for VGGT inference. Each submap contains 200 images and we set the confidence percentage to 55% for noise filtering of VGGT results (ignoring points with last 45% confidence).

3.1 Study Area

We use the dataset of Hong Kong medium, and Wuhan Loop 2 dataset of AGI2P dataset (Yang et al., 2026) for evaluation. The Hong Kong dataset (Hsu et al., 2023) is collected with a car and Wuhan dataset is acquired with a helmet-based mobile mapping system WHU-Helmet (Li et al., 2023). The resolution of the ALS point clouds is 0.5 m. Details of the images for the experiment are shown in Table 1.

Scene	Image number	Image resolution	Trajectory length (km)
Hong Kong	9041	672*376	3.64
Wuhan	24620	1280*720	3.43

Table 1. Dataset details for experiment.

Fig. 3 and Fig. 4 show the trajectories and global reconstructed image point clouds with ALS point clouds for Hong Kong and Wuhan datasets. ALS point clouds are rendered by height and red triangles are selected check points for accuracy evaluation.

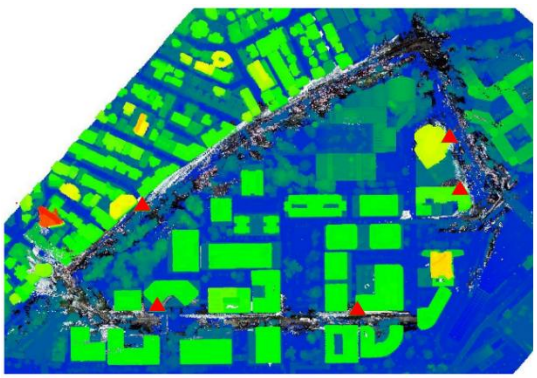

Figure 3. Point clouds visualization of Hong Kong.

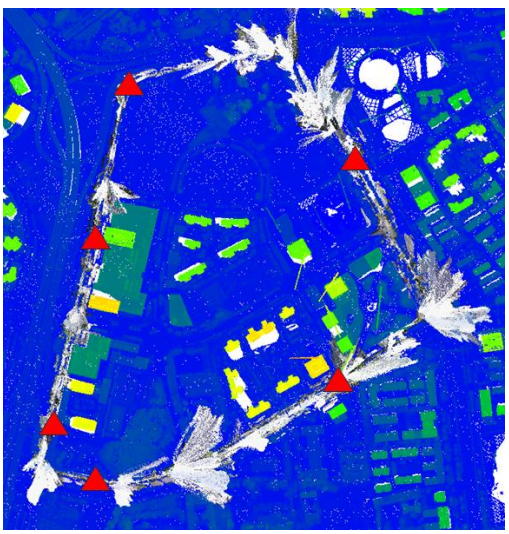

Figure 4. Point clouds visualization of Wuhan.

Figure 5. Raw images of Hongkong dataset in AGI2P.

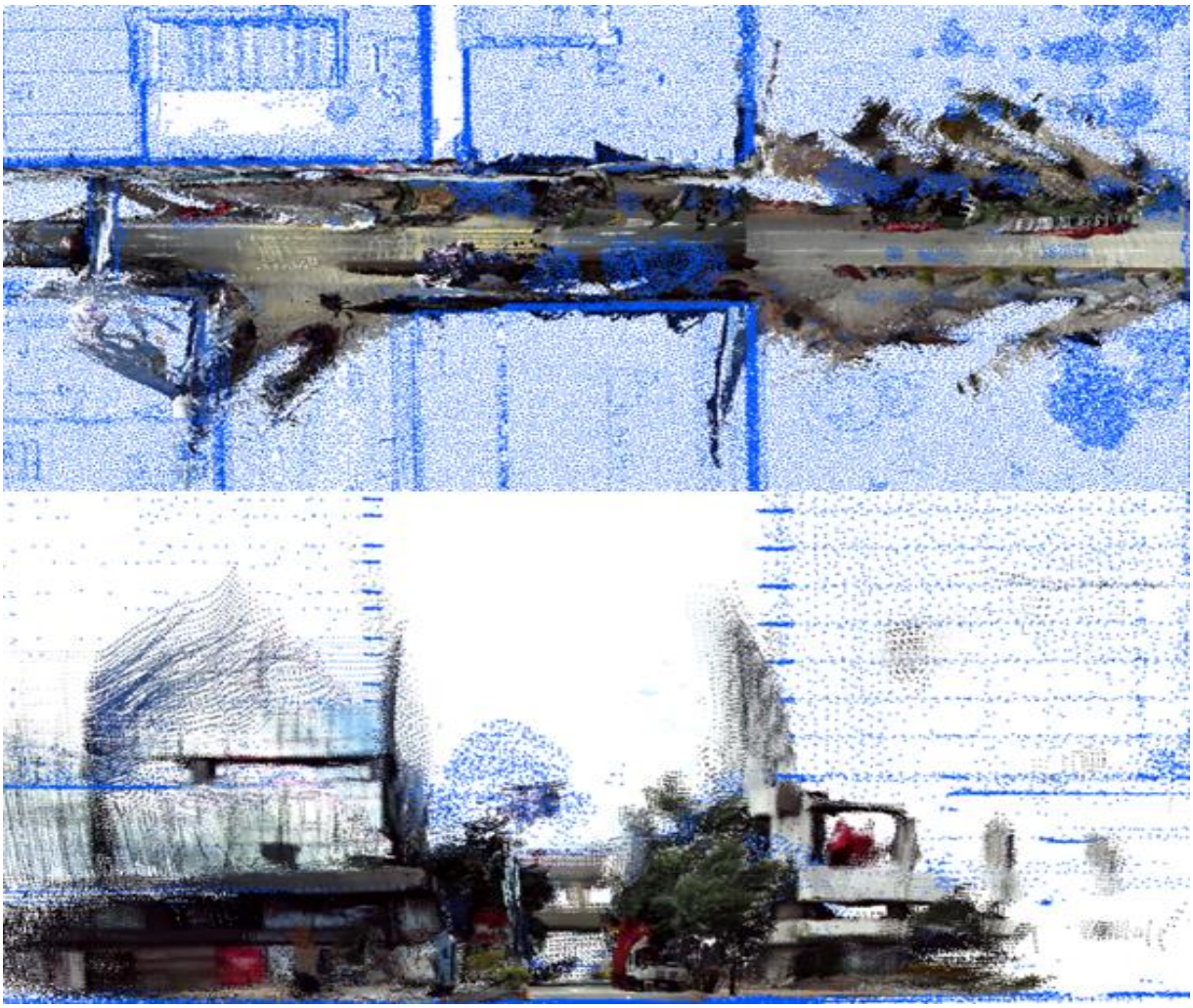

Figure 6. Details of image point clouds and ALS point clouds of Hongkong dataset in AGI2P.

Figure 7. Raw images of Wuhan dataset in AGI2P.

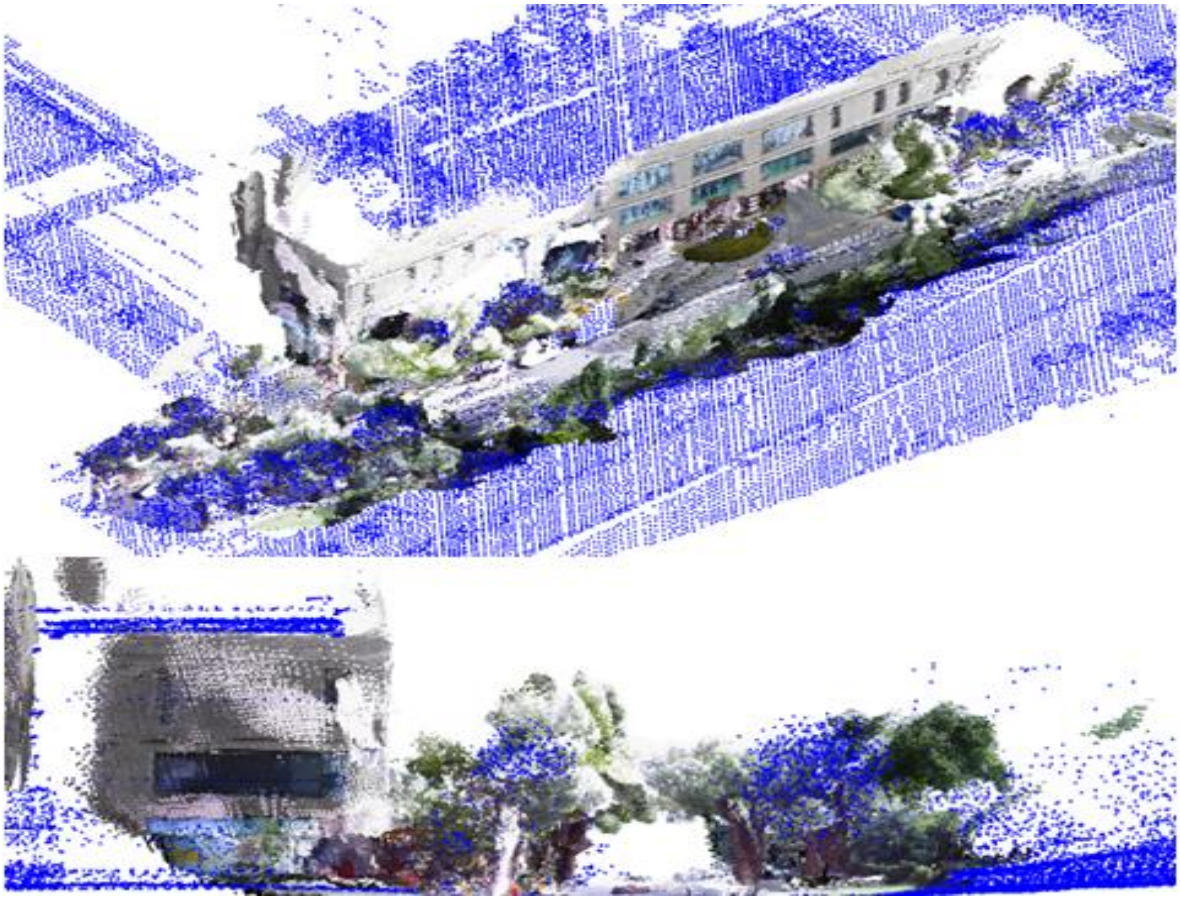

Figure 8. Details of image point clouds and ALS point clouds of Wuhan dataset in AGI2P.

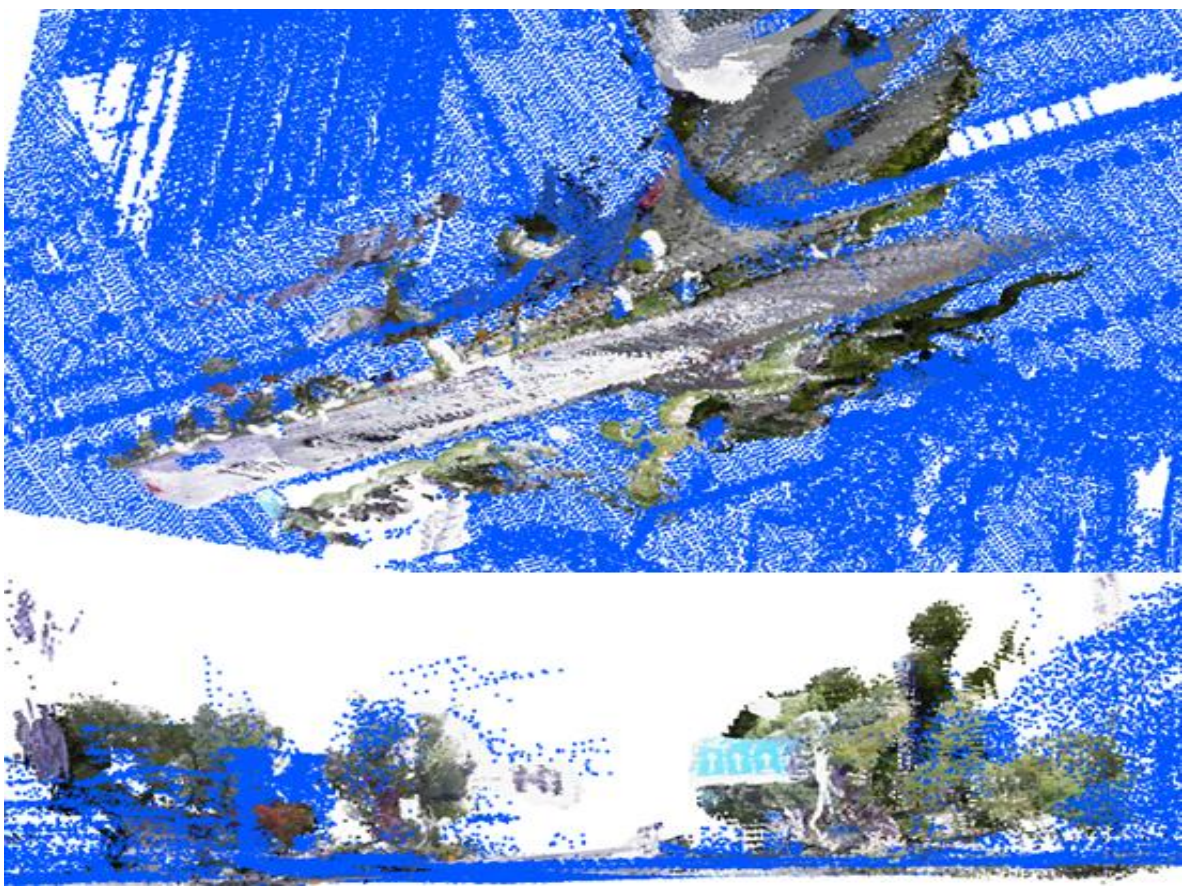

Figure 9. Details of image point clouds and ALS point clouds of Wuhan dataset in AGI2P.

https://doi.org/10.5194/isprs-archives-XLIX-B1-2026-77-2026

3.2 Accuracy Evaluation

To evaluate accuracy of the reconstructed image point clouds, we select 11 evenly distributed check points from ALS data and compute the offsets between corresponding points in image point clouds and ALS point clouds. The distribution of check points is demonstrated in Fig.3 and Fig. 4. Visualization results of image point clouds are demonstrated in Fig. 6, Fig. 8 and Fig. 9, where the corresponding input images are also shown in Fig. 5 and Fig. 7. Table 2 and Fig. 10 show the statics of check point errors.

Study Area	Check point #	Min offset	Max offset	Mean offset
Hong Kong	5	0.53	1.02	0.77
Wuhan	6	0.47	0.94	0.69

Table 2. Error distribution.

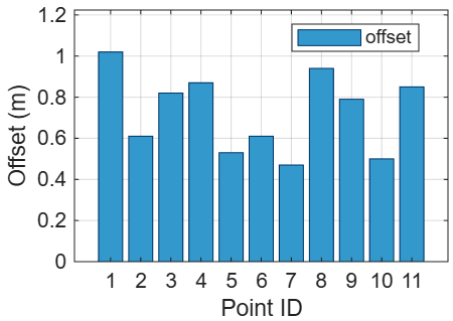

Figure 10. Check point offsets.

4. CONCLUSIONS

This paper proposes a method for applying VGGT with prior ALS point clouds in large-scale outdoor dense mapping tasks. The evaluation of check points accuracy demonstrates the average offsets of 0.77 m and 0.69 m in Hong Kong and Wuhan datasets, respectively, showing the capability of applying feed-forward models into large-scale 3D mapping. Future research will focus on extending the framework with multiple sensors.

REFERENCES

Chang, M.-F., Mangelson, J., Kaess, M., Lucey, S., 2021. Hypermap: Compressed 3d map for monocular camera registration, in: 2021 IEEE International Conference on Robotics and Automation (ICRA). pp. 11739–11745.

Deng, K., Ti, Z., Xu, J., Yang, J., Xie, J., 2025. VGGT-Long: Chunk it, Loop it, Align it–Pushing VGGT’s Limits on Kilometer-scale Long RGB Sequences. arXiv preprint arXiv:2507.16443.

El-Sheimy, N., Youssef, A., 2020. Inertial sensors technologies for navigation applications: State of the art and future trends. Satellite navigation 1, 2.

Hsu, L.-T., Huang, F., Ng, H.-F., Zhang, G., Zhong, Y., Bai, X., Wen, W., others, 2023. Hong Kong UrbanNav: An open-source multisensory dataset for benchmarking urban navigation algorithms. Navigation 70.

Kang, S., Liao, Y., Li, J., Liang, F., Li, Y., Zou, X., Li, F., Chen, X., Dong, Z., Yang, B., 2024. Cofii2p: Coarse-to-fine correspondences-based image to point cloud registration. IEEE Robot Autom Lett 9, 10264–10271.

Li, J., Wu, W., Yang, B., Zou, X., Yang, Y., Zhao, X., Dong, Z., 2023. WHU-helmet: A helmet-based multisensor SLAM dataset for the evaluation of real-time 3-D mapping in large-scale GNSS-denied environments. IEEE Transactions on Geoscience and Remote Sensing 61, 1–16.

Li, J., Xu, X., Liu, Z., Yuan, S., Cao, M., Xie, L., 2026. Aeos: Active environment-aware optimal scanning control for uav lidar-inertial odometry in complex scenes. ISPRS Journal of Photogrammetry and Remote Sensing 232, 476–491.

Li, J., Yuan, S., Cao, M., Nguyen, T.-M., Cao, K., Xie, L., 2024. HCTO: Optimality-aware LiDAR inertial odometry with hybrid continuous time optimization for compact wearable mapping system. ISPRS Journal of Photogrammetry and Remote Sensing 211, 228–243.

Maggio, D., Carlone, L., 2026. VGGT-SLAM 2.0: Real time Dense Feed-forward Scene Reconstruction. arXiv preprint arXiv:2601.19887.

Maggio, D., Lim, H., Carlone, L., 2025. Vggt-slam: Dense rgb slam optimized on the sl (4) manifold. arXiv preprint arXiv:2505.12549.

Murai, R., Dexheimer, E., Davison, A.J., 2025. Mast3r-slam: Real-time dense slam with 3d reconstruction priors, in: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 16695–16705.

Schwarz, K.P., El-Sheimy, N., 2004. Mobile mapping systems–state of the art and future trends. International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences 35, 10.

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

Wang, J., Karaev, N., Rupprecht, C., Novotny, D., 2024. Vggsfm: Visual geometry grounded deep structure from motion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21686–21697.

Wang, L., Guo, L., Xu, Z., Wang, Q., Gao, F., Chen, X., 2026. LiDAR-VGGT: Cross-Modal Coarse-to-Fine Fusion for Globally Consistent and Metric-Scale Dense Mapping. IEEE Robot Autom Lett 11, 4721–4728.

Xu, J., Qi, Z., Yan, Z., Gao, X., Jiao, Q., Xia, S., Chen, X., Pei, L., 2026. VGGT-MPR: VGGT-Enhanced Multimodal Place Recognition in Autonomous Driving Environments. arXiv preprint arXiv:2602.19735.

Yang, Y., Li, J., Liao, Y., Li, Y., Niu, R., Zhang, Y., Dong, Z., Yang, B., El-Sheimy, N., 2026. AGI2P: Benchmarking Aerial–Ground Image-to-Point cloud localization with a large-scale dataset. ISPRS Journal of Photogrammetry and Remote Sensing 236, 22–36.

Zheng, X., Wen, W., Hsu, L.-T., 2025. Tightly-Coupled Visual/Inertial/Map Integration With Observability Analysis for Reliable Localization of Intelligent Vehicles. IEEE Transactions on Intelligent Vehicles 10, 863–875.