The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume L-4/W2-2026
21st International 3D GeoInfo Conference 2026, 28–29 September 2026, Sofia, Bulgaria

Fine-Registration between Point Clouds and Aerial Images using Monocular Geometry

Qipeng Mei[1], Dimitri Bulatov[2], Dorota Iwaszczuk[1]

[1] Technical University of Darmstadt, Germany - (qipeng.mei,dorota.iwaszczuk)@tu-darmstadt.de
[2] Fraunhofer IOSB, Ettlingen, Germany - dimitri.bulatov@iosb.fraunhofer.de

Keywords: LiDAR–image registration, cross-modal, building, depth estimation, planar roofs.

Abstract

Accurate registration between aerial imagery and LiDAR point clouds is fundamental to building-level analysis and urban modeling. Although coarse alignment can be achieved through geo-referencing or sensor calibration, residual misalignment often remains and affects the reliable interpretation of roof structures. To address this issue, we propose a Progressive Correlation-based Geometric Registration (PCGR) framework for fine registration between projected LiDAR observations and aerial images using monocular geometry. We leverage MoGe-2, a state-of-the-art tool for depth synthesis from optical images, trained on large-scale data to recover a dense depth map from a single aerial patch around the building. Such models encode strong geometric priors and provide a geometrically consistent representation for building structures. Sparse LiDAR points are aligned with the monocular depth through a progressive coarse-to-fine optimization strategy. The alignment is guided by the Pearson correlation coefficient, ensuring robustness to scale and bias differences. We evaluate the method on those images, exhibit significant deviations from the LiDAR point cloud to correct them for our ongoing task, namely, roof detail analysis in 3D. Experimental results demonstrate consistent improvements in alignment quality, including a systematic reduction in height standard deviation and decreased facet-wise RMS residuals in most cases. These findings indicate that large-scale learned monocular geometry effectively bridges the representation gap between aerial imagery and LiDAR for building-level cross-modal registration.

1. Introduction

Buildings represent an essential part of urban infrastructure and their correct geo-localization is important for many applications, such as city planning, climate science, civil security, and disaster management (Bulatov, 2019). As a consequence, increasing multi-sensor coverage of buildings is required and is actually being implemented. This in turn, requires developing advanced methodology on fusion of data produced by different sensors, such as airborne and terrestrial laser point clouds, satellite and drone imagery, GIS data and in-situ sensors (Shahraki et al., 2024; Guo et al., 2025). As Haklay (2010); Fan et al. (2014) have pointed out, the deviations between the GIS database and the image can be quite high ($\approx$ 8 m) and the main reasons for this are: the three-dimensional character of buildings, occasional discrepancy between roof polygons and ground plans, as well as changes with respect to the out-to-date database. However, to perform the analysis of roof structures for the aforementioned applications, improved registration is crucial.

In practice, this problem is commonly encountered under an almost-nadir imaging setup, where a coarse alignment between modalities is already available through geo-referencing or sensor calibration. The remaining challenge is therefore not global registration, but the refinement of residual misalignment at the building level. This setting imposes specific constraints, as the misalignment is typically small but structurally significant, especially for complex roof geometries. A key difficulty lies in the heterogeneous nature of the data. Aerial RGB imagery and LiDAR observations differ substantially in representation, making direct correspondence estimation unreliable. In particular, LiDAR data is often sparse, for example, for water and tar, and unevenly distributed, while RGB images lack explicit geometric information. As a result, conventional registration approaches based on photometric or geometric similarity are not well-suited for this cross-modal scenario.

Recent advances in monocular geometry estimation provide a promising direction to address this challenge. By transforming RGB imagery into dense depth representations, these methods enable the extraction of complete geometric structures from a single image. This allows the registration problem to be reformulated as a geometry-consistent alignment task, where dense monocular predictions serve as structural priors to guide the refinement of sparse LiDAR observations.

In this paper, we propose a Progressive Correlation-based Geometric Registration (PCGR) framework to refine the alignment between projected LiDAR points and monocular depth predictions within building regions. The proposed method formulates the problem as a correlation-driven geometric alignment process and solves it using a progressive coarse-to-fine optimization strategy, enabling robust initialization and accurate refinement under small misalignment conditions.

We evaluate the proposed approach on multiple building scenes with varying structural complexity. Experimental results demonstrate consistent improvements in geometric alignment, particularly in scenarios with sparse LiDAR observations and complex roof structures.

Our contributions can be summarized as follows:

- We formulate building-level cross-modal registration as a geometry-consistent alignment problem under an almost-nadir setting.
- We leverage monocular geometry estimation to bridge the representation gap between RGB imagery and LiDAR data.
- We propose a progressive correlation-based optimization framework for robust and accurate alignment refinement.

- We validate the effectiveness of the method through both qualitative and quantitative evaluations on diverse building scenarios.

2. Related Work

Aerial optical-to-LiDAR registration has been extensively investigated in the context of 3D city modeling, where reliable alignment enables consistent reconstruction of building facades/roofs and accurate estimation of geometric attributes (Abayowa et al., 2015). To bridge the 2D image domain and 3D point clouds, a common strategy is to introduce intermediate geometric representations and establish correspondences using structural primitives that are comparatively stable across modalities. Tamas and Kato (Tamas and Kato, 2013) formulate targetless calibration as a 2D–3D registration problem based on planar regions observable in both modalities. LiDAR point clouds can also be aligned to optical imagery by matching line and point features, exploiting the fact that edges and man-made structures provide strong geometric cues even when radiometric characteristics differ (Lv and Ren, 2015). In addition to such geometry-based correspondence strategies, some methods transform LiDAR data into image-like representations (e.g., intensity images) to enable appearance-based matching, where feature correspondences and mutual information are jointly exploited for 2D–3D registration (Bodensteiner and Arens, 2012).

Beyond explicit feature matching, several works formulate registration as a geometry-driven optimization problem. Huang et al. (2018) propose a method based on the closest point principle and collinearity equations, which directly minimizes the distance between photogrammetric image points and LiDAR surfaces without requiring explicit extraction of linear or planar features. This formulation reduces the dependency on handcrafted feature detection and improves robustness in large-scale aerial scenarios.

Furthermore, hierarchical strategies have been explored to improve robustness in complex urban environments. Coarse-to-fine approaches first establish a global alignment using structural elements such as buildings and then refine the transformation via local optimization or similarity measures, enabling more stable convergence under large initial misalignment (Lee et al., 2024).

More recently, learning-based approaches have been introduced to improve the robustness of cross-modal correspondence estimation. Fervers et al. (2022) propose a geo-tracking framework that leverages metric learning to extract features from ground and aerial images, and projects ground features into a bird's-eye-view representation using LiDAR to facilitate cross-view matching. Instead of relying solely on geometric primitives, Leahy and Jabari (2024) enhances LiDAR representations by constructing multi-channel feature layers from elevation, intensity, and bearing angle attributes, and employs a graph neural network (GNN) to establish correspondences between optical images and LiDAR-derived representations. The final alignment is then estimated via a 2D–3D affine transformation.

With recent advances in monocular depth estimation, it has become possible to recover dense and reliable geometric structure from RGB images, providing an effective bridge for cross-modal registration with LiDAR. In the autonomous driving domain, several works exploit monocular depth as an intermediate representation for alignment. Borer et al. (2024) employs monocular depth within a mutual information maximization framework to perform geometry-driven alignment without supervised training. In contrast, DF-Calib (Han et al., 2025) formulates calibration as a depth-to-depth flow estimation problem based on dense depth prediction and completion, enabling learned dense correspondences.

3. Methodology

As illustrated in Figure 1, we propose a Progressive Correlation-based Geometric Registration (PCGR) framework to refine the alignment between projected LiDAR observations and monocular depth predictions within building regions. Unlike general-purpose registration, we focus on an almost-nadir setting with coarse pre-alignment, where residual misalignment is small.

Given an input image, a monocular model produces a dense depth map, providing a geometric prior for building structures, while sparse LiDAR points are projected onto the image plane. Based on this, PCGR formulates the task as a correlation-driven geometric alignment problem and solves it using a progressive coarse-to-fine optimization strategy for robust and accurate refinement.

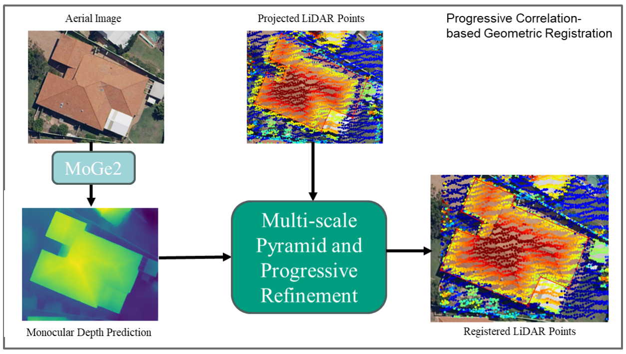


Figure 1. Pipeline of Progressive Correlation-based Geometric Registration.

3.1 Accurate Monocular Geometry Estimation

We employ MoGe-2 Wang et al. (2025b), an advanced monocular geometry estimation model capable of recovering a metric-scale depth map from a single image. MoGe-2 extends the prior MoGe framework Wang et al. (2025a), which predicts affine-invariant geometry with unknown scale, by introducing strategies for metric-scale reconstruction while preserving accurate relative geometry. Unlike conventional monocular depth estimation methods that rely on depth regression, MoGe-2 adopts a geometry-centric formulation and directly predicts a dense 3D point map of the scene. This representation provides a more straightforward description of 3D geometry compared to depth-based approaches. Furthermore, it alleviates ambiguity in geometry reconstruction, since depth-based methods require known camera intrinsics to recover 3D structure. By extending the affine-invariant representation to metric geometry prediction, the model is able to recover both consistent global scale and accurate structural relationships.

In addition, MoGe-2 introduces a unified data refinement strategy to address noise and inconsistencies in real-world training data. By improving the quality of supervision, the model enhances the reconstruction of fine-grained geometric details.

Combined with large-scale training on diverse datasets, MoGe-2 demonstrates strong generalization across different scenes while simultaneously providing accurate relative geometry, reliable metric scale, and detailed structural reconstruction.

These properties make MoGe-2 particularly suitable for our cross-modal alignment task, where the predicted geometry serves as a dense and geometrically consistent reference to guide the registration of sparse LiDAR projections.

3.2 Projection and Sparse Depth Grid Construction

Because the image patch is quasi-nadir and nearly georeferenced, the UTM coordinates of the LiDAR points can be transferred into the coordinate system of the image patch, $X = (x, y, z)$, where x, y are the pixel coordinates and z is the elevation. Thus, the projected points are rasterized into a sparse depth grid $Z \in \mathbb{R}^{H\times W}$, along with a corresponding validity mask $M \in \{0, 1\}^{H\times W}$, where depth values are assigned only to pixels with valid projections. The similarity computation is restricted to the set of valid overlapping pixels:

$$\Omega = \{i \mid M_i = 1\}. \tag{1}$$

To achieve fine registration, the rigid transformation T is applied to x and y coordinates.

3.3 Multi-scale Pyramid and Progressive Refinement.

We model the 2D alignment between D and Z using a metric transformation parameterized by $\boldsymbol{\xi} = [d_x, d_y, \theta]$, which consists of translations (d_x, d_y) and a rotation θ. Given a point (x, y), the transformed coordinates are defined as

$$T(\boldsymbol{\xi}) = \begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} d_x \\ d_y \end{bmatrix}. \tag{2}$$

In this work, we primarily focus on estimating the translation (d_x, d_y) in a coarse-to-fine manner, while the rotation θ is refined only at the final stage.

To avoid exhaustive high-dimensional search while maintaining both accuracy and computational efficiency, we construct a multi-scale pyramid representation $\{D^s, Z^s, M^s\}$, where s denotes the scale level. A coarse-to-fine strategy is adopted for efficient optimization.

At the coarsest level $s = S$, a large search range is used to enumerate candidate translations (d_x, d_y), and the corresponding score ρ_s is evaluated. The optimal solution (d_x^*, d_y^*) is obtained by selecting the translation with the highest score (arg-max) and then propagated to the next finer level by scale adjustment:

$$(d_x(s-1)^0, d_y(s-1)^0) = 2\,(d_x(s)^*, d_y(s)^*). \tag{3}$$

At each subsequent level $s - 1$, the search is restricted to a small neighborhood around the propagated initialization $(d_x(s-1)^0, d_y(s-1)^0)$, and the optimal translation is likewise obtained by exhaustively evaluating ρ_{s-1} over the neighborhood and selecting the arg-max. This process continues until the finest level $s = 0$ is reached, whose solution is taken as the final translation estimate.

In other words, the reduction factor in the number of operations is $2^2 = 4$. The number of pyramid levels is not chosen arbitrarily, but is implicitly related to the ground sampling distance (GSD) of the LiDAR point cloud and the spatial resolution of the image. Specifically, the coarsest level should be defined such that the initial misalignment is within a manageable search range, while the finest level reflects the resolution at which geometric details are to be aligned. If too many levels are introduced, the resolution at coarse scales becomes overly reduced, which may lead to the loss of meaningful geometric structures and unnecessary computational overhead. Conversely, if too few levels are used, the optimization may fail to handle large initial misalignments, as the search at finer resolutions becomes prone to local minima. Therefore, the number of levels should be selected to balance robustness against large displacements and the preservation of fine-scale geometric details.

Finally, a local search over the rotation parameter θ is performed at the finest scale, optionally allowing small adjustments to (d_x, d_y).

3.4 Correlation-based Similarity Score

To supervise the alignment process, we adopt the Pearson correlation coefficient as the similarity objective.

Our formulation is based on the assumption that, under correct alignment, the depth values from LiDAR projections and monocular predictions follow an approximately linear relationship over the valid overlapping region Ω. This assumption is justified by the fact that both modalities observe the same underlying scene geometry, up to a potential scale and bias difference.

Specifically, the similarity score at scale s is defined as

$$\rho_s = \frac{\sum_{i\in\Omega} \left(Z_i^s - \bar{Z}^s\right)\left(D_i^s - \bar{D}^s\right)}{\sqrt{\sum_{i\in\Omega} \left(Z_i^s - \bar{Z}^s\right)^2 \sum_{i\in\Omega} \left(D_i^s - \bar{D}^s\right)^2}}, \tag{4}$$

where $\bar{Z}^s$ and $\bar{D}^s$ denote the mean values over the valid region Ω, and Z_i^s represents the translated sparse depth.

The Pearson correlation coefficient offers several advantages in our setting. First, it is invariant to affine transformations of depth values, making it robust to scale differences and bias between LiDAR and monocular predictions. Second, it captures global linear consistency rather than relying on point-wise absolute differences, which improves robustness to noise and outliers in sparse observations. Finally, it provides a bounded and well-behaved objective, facilitating stable optimization across pyramid levels.

4. Experiments

Compared to other types of objects, buildings exhibit regular geometric structures and clear boundaries, making them well-suited for evaluating geometric alignment. Furthermore, their correspondence in elevation is perceptually distinguishable, allowing alignment quality to be reliably assessed through direct visual inspection. In addition, buildings typically undergo minimal changes over time, providing a stable reference for consistent evaluation across different observations. Based on these properties, we construct our evaluation dataset using building instances extracted from aerial imagery and corresponding LiDAR observations.

https://doi.org/10.5194/isprs-archives-L-4-W2-2026-125-2026 | © Author(s) 2026. CC BY 4.0 License.

4.1 Dataset

City of Melville, a local council in Perth, Australia, acquired in February 2016 aerial datasets to perform various studies closely related to the urban heat island effect Council of City of Meville (2017). Such studies include determining the correlation between the heat and land usage, analyzing the shading and cooling properties of both native and exotic tree species, as well as localizing the most commonly used roofing materials. Among others, the LiDAR point cloud was acquired from an airborne system, using the scanner Riegl VZ - 1000. It has the density of approximately 2 - 4 points per square meter and an accuracy of 0.1 m. A high-resolution RGB aerial imagery dataset that was acquired in May of 2016 with a GSD of 0.1 m was provided by the company Spookfish, now Aerometrex. A shapefile containing the building outlines was available, as well. It was originally dated in 2012 and improved using Machine-Learning-based classification with NDSM and multispectral orthophoto. For details on data preparation, we refer to (Ilehag et al., 2018; Bulatov et al., 2020).

Unfortunately, there were still some registration errors in the building positions of the corrected shapefile and, besides, georefenciation errors in Spookfish images. These errors are mostly small, below 5 pixels, but are still considerable when it comes to the analysis of roof structures for building modeling with automatic methods. We computed a patch around each building and clipped the LiDAR point in the way shown in our pipeline in Figure 1. In addition, we manually annotate building outlines as well as planar structures within each building. These annotations provide accurate geometric references and are used for both qualitative visualization and quantitative evaluation of registration performance. In total, the dataset consists of 23 building instances used for evaluation.

4.2 Experimental Setup

We employ a three-level image pyramid to estimate the optimal alignment parameters in a coarse-to-fine manner. The pyramid is constructed with a scaling factor of 2 between adjacent levels, enabling efficient exploration of large search spaces while preserving fine-grained accuracy.

At the coarsest level, a global search is performed over translation parameters (d_x, d_y) within the range of $[-8, 8]$ pixels, which corresponds to $[-24, 24]$ pixels in the original image resolution. This allows the method to handle large initial misalignment. At the intermediate level, the search range is reduced to $[-4, 4]$ pixels (i.e., $[-8, 8]$ in the original scale), focusing on refining the alignment around the coarse estimate. Finally, at the finest level, a local search within $[-2, 2]$ pixels is conducted for fine-grained refinement. Rotation is only optimized at the finest level to avoid exhaustive high-dimensional search. Specifically, the rotation angle θ is searched within a range of $[-5^\circ, 5^\circ]$ with a step size of 0.25°. This design allows for accurate estimation of in-plane rotation while maintaining computational efficiency. Overall, the proposed coarse-to-fine strategy enables an effective search range of approximately $[-34, 34]$ pixels in the original image resolution, while maintaining computational efficiency through progressive refinement.

For comparison, a brute-force search at the original resolution over $d_x, d_y \in [-34, 34]$ and $\theta \in [-5^\circ, 5^\circ]$ with a step size of 0.25° would require evaluating

$$69 \times 69 \times 41 = 195{,}201$$

candidate transformations. In contrast, the proposed coarse-to-fine strategy reduces the search to only a small subset of candidates across pyramid levels. Specifically, the translation search requires 17×17, 9×9, and 5×5 evaluations from coarse to fine, while rotation is only optimized at the finest level over 41 angle candidates. As a result, the total number of evaluations is reduced to 1,395 in the joint local search setting, or 436 when translation and rotation are decoupled, yielding approximately $140\times$ to $448\times$ fewer evaluations than brute-force search.

4.3 Quantitative Evaluation Metrics

4.3.1 Height Standard Deviation

To quantitatively evaluate the alignment quality, we adopt the Height Standard Deviation (H-Std) within building regions as a measure of geometric consistency. Given that building outlines are manually annotated, we assume that, under accurate alignment, LiDAR points falling inside the outline predominantly belong to the building surface.

Let $\mathcal{B}$ denote the set of pixels within the annotated building footprint. For each valid projected LiDAR point $i \in \mathcal{B}$, we consider its height value z_i. Ideally, after correct registration, these points should lie on relatively smooth roof surfaces, resulting in reduced height variation. In contrast, misaligned points or points from other structures (e.g., ground or surrounding objects) introduce larger height discrepancies, thereby increasing the standard deviation.

To improve robustness against extreme values such as roof ridges, we exclude the top 5% highest points based on height. Specifically, let $\mathcal{B}'$ denote the filtered set obtained by removing points whose height exceeds the 95th percentile. The H-Std is then computed as:

$$\sigma_{\text{H-Std}} = \sqrt{\frac{1}{|\mathcal{B}'|} \sum_{i \in \mathcal{B}'} (Z_i - \bar{Z})^2}, \tag{5}$$

where $\bar{Z}$ is the mean height over $\mathcal{B}'$.

A lower value of H-Std indicates better alignment quality, as it reflects reduced height variation and fewer outlier points introduced by misalignment.

4.3.2 Root Mean Square Residual

To further evaluate geometric consistency at a finer level, we adopt the root mean square (RMS) residual of point-to-plane distances as a measure of surface flatness. Here, we leverage manually annotated roof structures and compute flatness separately for each roof facet.

The RMS residual for each facet is defined as:

$$\text{RMS} = \frac{1}{K} \sum_k \left(\frac{1}{\#X_k} \sum_{X \in B_k} (\pi_k^T X)^2 \right)^{\frac{1}{2}}, \tag{6}$$

where K denotes the total number of planar regions, and k indexes each individual region. Moreover, B_k represents the set of 3D points belonging to the k-th region, $\#X_k$ denotes the number of points in B_k, and $X \in \mathbb{R}^4$ is a 3D point expressed in homogeneous coordinates. Finally, $\pi_k \in \mathbb{R}^4$ denotes the plane parameters of the k-th region, computed using direct linear transformation, so that $\pi_k^T X$ is the algebraic distance from point X to the corresponding plane.

https://doi.org/10.5194/isprs-archives-L-4-W2-2026-125-2026

This facet-wise formulation provides a more precise evaluation of alignment quality, as points are expected to lie on their corresponding planar surfaces. Accurate registration leads to smaller deviations within each facet, whereas misalignment or contamination from non-planar structures results in increased residuals. Therefore, a lower RMS indicates better alignment and improved geometric consistency at the structural level.

5. Evaluation

5.1 Qualitative Evaluation

Figure 2 presents qualitative results of LiDAR point cloud registration on four building scenes. For each example, the top-left shows the aerial image with building outlines, while the top-right displays the MoGe-2 predictions overlaid with the same outlines. The bottom-left illustrates the misaligned LiDAR projections overlaid on the aerial image, and the bottom-right shows the corresponding results after registration. The optimal registration parameters are also shown in an inset box.

Figure 2(a) illustrates the case of a simple roof structure, where the LiDAR points are effectively aligned with the underlying geometry after registration. Notably, the elevation difference between the left and right parts of the roof is clearly preserved, which is also well captured by the monocular depth prediction, providing reliable guidance for alignment.

Figure 2(b) presents a complex single-layer roof structure with irregular boundaries. Despite the increased geometric complexity, the proposed method successfully corrects spatial misalignment and brings the LiDAR points into close correspondence with the building footprint, while preserving the overall structure.

Figure 2(c) shows a complex multi-layer roof structure with multiple elevation levels. The initial LiDAR projections exhibit significant misalignment across different roof components, whereas the proposed method effectively aligns points belonging to different layers, restoring structural coherence.

Figure 2(d) illustrates a challenging scenario involving low-reflectance regions, where LiDAR returns are sparse due to material properties. Even under such degraded sensing conditions, the method is able to recover a reasonable alignment, demonstrating robustness by leveraging the dense geometric guidance from monocular prediction.

5.2 Quantitative Evaluation

Beyond the qualitative evaluation, we further conduct a quantitative analysis to assess the registration performance from a geometric perspective.

Figure 3 presents the comparison of the Pearson Correlation Coefficient (PCC) before and after registration. Since PCC is directly adopted as the optimization objective in our framework, an increase in PCC for a given sample is expected by construction. Therefore, this plot confirms that the optimization converges successfully and consistently across all evaluated buildings, rather than failing or diverging. More specifically, as shown in the figure, all samples lie above the identity line $y = x$, indicating that the PCC consistently increases after registration for every evaluated building. Moreover, a considerable number of samples are distributed closer to the upper region of the plot, with several approaching the $y = 2x$ reference line, reflecting cases where the pre-registration correlation was particularly low and the optimization had greater room for improvement.

Then, Figures 4 and 5 present the before-and-after comparisons of H-Std and RMS, respectively. In both figures, the dashed identity line $y = x$ denotes unchanged values before and after registration, while the additional auxiliary dashed lines indicate fixed multiplicative ratios between the two measurements, providing a visual reference for the magnitude of improvement or degradation. Points below the identity line correspond to reduced metric values after registration, whereas points above it indicate increased values.

Figure 4 shows the comparison of H-Std before and after registration. All samples are located below the identity line $y = x$, indicating that H-Std is reduced for every evaluated building after registration. Moreover, several points lie close to the $y = 0.5x$ line, suggesting that the reduction approaches a two-fold improvement for some buildings. Overall, the distribution shows a consistent decrease in H-Std across all samples, with improvements ranging from moderate reductions to cases approaching a factor of two.

Figure 5 shows the comparison of RMS before and after registration. Here, each point represents the mean RMS residual of all annotated roof planes within one building. Most samples are distributed below the identity line $y = x$, indicating that the mean RMS residual decreases after registration for the majority of evaluated buildings. Specifically, 82.6% of the buildings exhibit lower RMS values after registration. Several points lie close to the $y = 0.5x$ line, and a few even fall below it, suggesting nearly two-fold or greater improvement in some cases. Meanwhile, a limited number of samples are located above the identity line, indicating increased RMS residuals for a small subset of buildings.

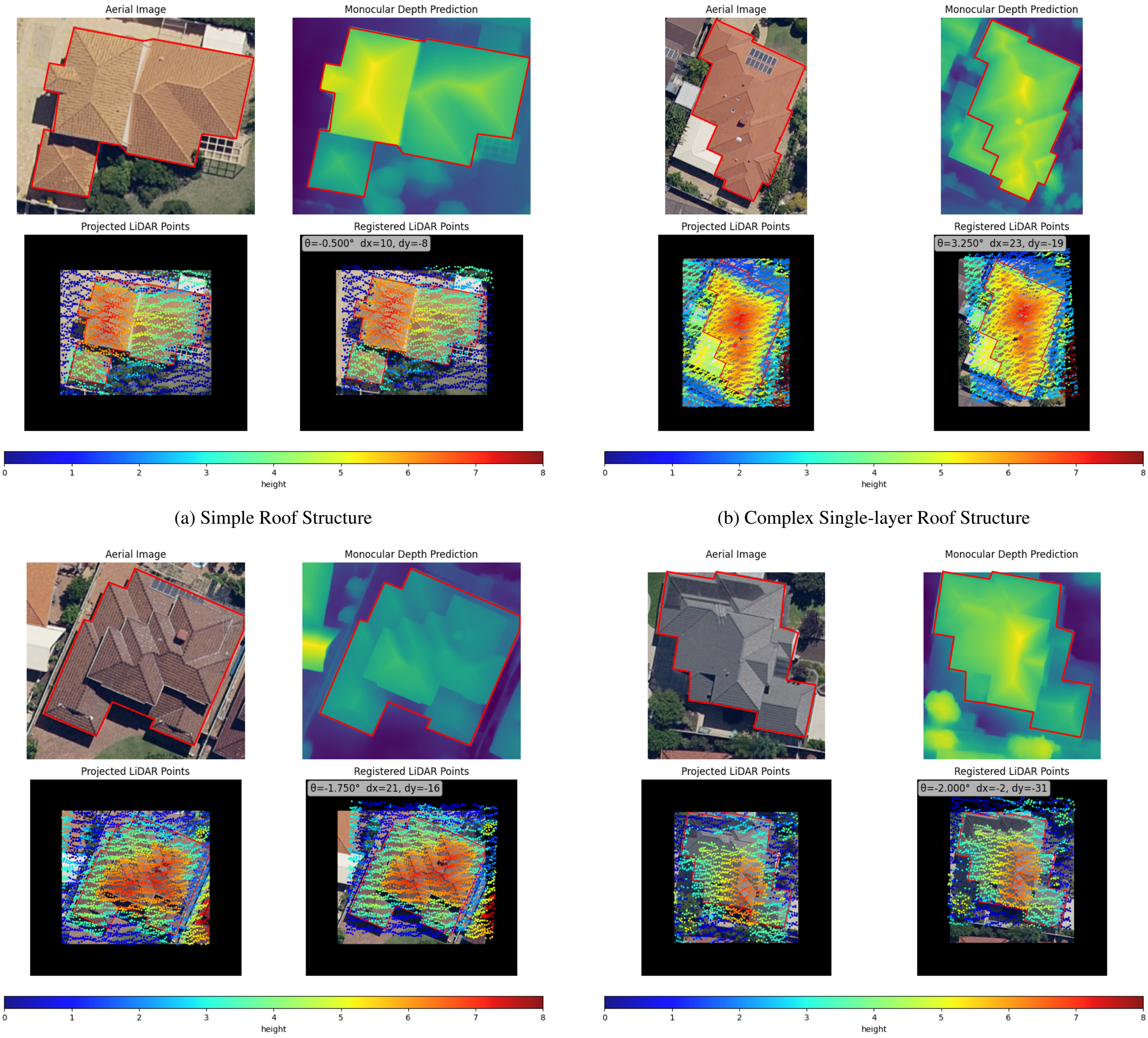
6. Discussion

The qualitative and quantitative results consistently demonstrate that the our method improves cross-modal alignment between LiDAR observations and monocular depth predictions on building level.

The qualitative results reveal that the method generalizes well to buildings with varying structural complexity, including multi-layer roofs and irregular geometries. Even in challenging scenarios with sparse LiDAR returns (e.g., low-reflectance regions), the method is able to recover plausible alignment by relying on dense monocular predictions. This highlights the advantage of cross-modal fusion, where complementary sensing modalities compensate for each other's weaknesses.

The consistent reduction in H-Std across all evaluated buildings indicates that the proposed method successfully suppresses height inconsistencies within building regions. Specifically, MoGe-2 provides dense and geometrically consistent depth predictions that serve as a reliable global reference, while sparse LiDAR observations offer accurate but localized geometric constraints. The use of Pearson correlation further enables a robust measurement of structural consistency between the two representations, allowing the method to effectively align sparse observations with dense predictions.

https://doi.org/10.5194/isprs-archives-L-4-W2-2026-125-2026



(a) Simple Roof Structure

(b) Complex Single-layer Roof Structure

(c) Complex Multi-layer Roof Structure

(d) Complex Roof Structure with Low-reflectance Regions

Figure 2. Qualitative results of LiDAR point cloud registration across different roof structures. Color indicates relative height (m).

To enable more detailed observations, we further adopt the RMS metric. In contrast to H-Std, which captures global height variation within building footprints, RMS focuses on local surface consistency by quantifying point-to-plane residuals at the roof-facet level. The results indicate that, beyond reducing global height variation, the proposed method effectively enhances fine-grained geometric consistency at the roof-facet level. This implies that our method not only removes points that do not belong to the roof surface, such as ground or surrounding objects, but also enforces consistency among different roof facets. Notably, while H-Std shows uniform improvement across all samples, RMS exhibits a small number of degraded cases, indicating that local surface fitting is more sensitive to residual noise and imperfect depth estimation. This apparent degradation is mainly caused by outliers on rooftop surfaces, such as chimneys, antennas, or overhead transmission lines. These structures violate the planar assumption used in RMS computation and can significantly influence the residual error. As a result, even if the overall geometric alignment is improved, the presence of such non-planar elements may lead to higher RMS values in certain cases. Furthermore, the planar assumption inherently treats these rooftop objects as outliers during evaluation. Nevertheless, the overall distribution clearly demonstrates that the proposed method consistently improves geometric alignment for the majority of buildings. The substantial reduction in RMS values across most samples indicates that the method effectively enhances the alignment quality despite the presence of local structural disturbances.

Despite the encouraging results, several limitations remain. The proposed framework has been evaluated on a limited number of building instances, which may not fully capture the variability of real-world conditions. In particular, the current dataset primarily consists of buildings with complex roof structures, while more regular roof types—such as flat roofs, gable roofs, and shed roofs—have not yet been included. Although these simpler geometries may provide fewer distinctive geometric features for registration due to their more regular and continuous surfaces, it is hypothesized that the height differences between these roofs and their surrounding ground objects, or

https://doi.org/10.5194/isprs-archives-L-4-W2-2026-125-2026 | © Author(s) 2026. CC BY 4.0 License.

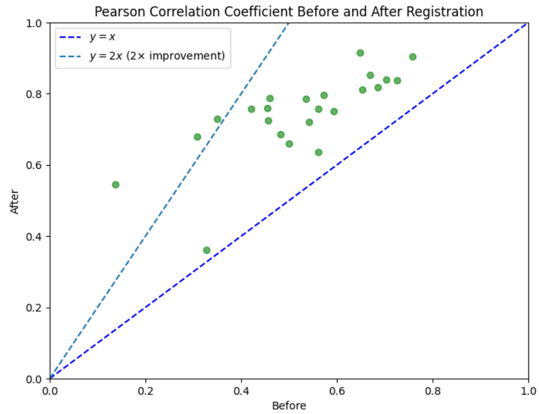


Figure 3. Comparison of PCC before and after registration.

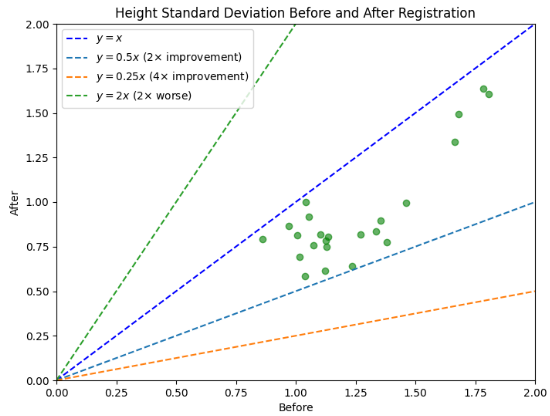


Figure 4. Comparison of H-Std before and after registration.

the surrounding terrain more broadly, could serve as auxiliary cues to assist PCGR in achieving robust registration.

Beyond the limited diversity of roof types, the overall registration quality is also inherently tied to the accuracy of the monocular depth predictions, since the proposed framework relies on MoGe-2 as its source of dense depth estimates. Although the current experiments did not reveal severe failure cases attributable to depth estimation errors, a more systematic analysis of this sensitivity—for instance, by introducing controlled perturbations to the depth predictions or by evaluating performance under varying imagery quality—would provide deeper insight into the robustness of the proposed framework and is left for future investigation.

In addition to these methodological considerations, the current evaluation primarily focuses on quantifying the improvement achieved by the proposed optimization relative to the initial misalignment, rather than benchmarking against established registration approaches. While this evaluation protocol effectively demonstrates the internal consistency and effectiveness of the proposed method, it does not provide a direct measure of its relative competitiveness compared to existing techniques, such as edge-based or feature-based registration methods. A comprehensive comparative study would offer a more complete understanding of the advantages and limitations of the proposed framework relative to the state of the art, and is planned as part of future work.

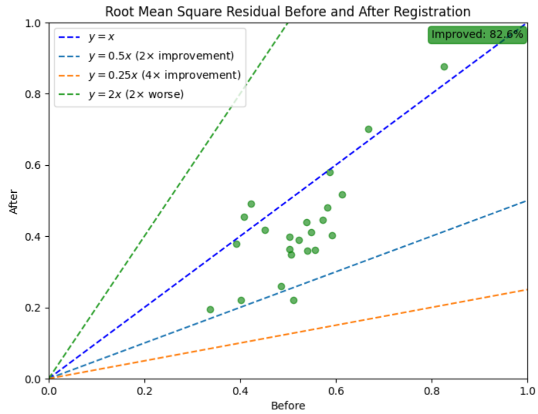


Figure 5. Comparison of RMS before and after registration. Each point represents the mean RMS residual of all annotated roof planes within one building.

7. Conclusion

We presented a Progressive Correlation-based Geometric Registration (PCGR) framework for fine registration between aerial images and LiDAR point clouds at the building level. The key idea is to use monocular geometry estimation to convert aerial RGB imagery into a dense geometric representation, which serves as a structural reference for aligning sparse LiDAR projections. Benefiting from recent advances in large-scale monocular geometry learning, the adopted MoGe-2 model provides high-quality and geometrically consistent depth predictions, even in complex urban scenes. This strong geometric prior plays a crucial role in bridging the representation gap between image and point cloud modalities. Based on this representation, the registration task is formulated as a correlation-driven geometric alignment problem and solved with a progressive coarse-to-fine optimization strategy.

The experimental results demonstrate that the proposed method can effectively refine residual misalignment across buildings with different roof structures. This improvement is evident from qualitative results, which show better alignment between projected LiDAR points and building geometry, even in challenging scenarios such as multi-layer roofs and sparse LiDAR observations. These observations are further supported by quantitative results. The quantitative results show consistent reductions in H-Std within building regions. Facet-wise RMS residuals decrease in the majority of cases. These results suggest that our proposed framework is effective not only in reducing global geometric inconsistency but also in improving local structural alignment.

Future work will focus on extending the evaluation to more diverse datasets, including a wider range of roof types and more challenging acquisition conditions such as low-resolution imagery, non-nadir viewing geometries, and large initial misalignments involving significant rotations. In addition, future efforts will investigate the sensitivity of the registration to depth prediction errors and conduct comparative evaluations against established registration approaches, in order to further validate the robustness and competitiveness of the proposed framework.

The proposed framework demonstrates promising performance in refining cross-modal alignment thanks to the tremendous advance in style-transfer techniques in general and in monocular geometry learning in particular. In this regard, it would be interesting to investigate the extension of the proposed framework to other sensing modalities, such as synthetic aperture radar (SAR) imagery, where substantial appearance differences further increase the difficulty of cross-modal alignment.

Beyond these aspects, the proposed method provides a solid foundation for downstream applications. In particular, the improved alignment between aerial imagery and LiDAR data enables more reliable and detailed 3D roof structure analysis, where accurate geometric consistency is essential for capturing fine-scale structural characteristics. Future work will therefore focus on leveraging the proposed registration framework to support reconstruction and interpretation of complex roof structures in urban environments.

Acknowledgement

The authors thank the China Scholarship Council (CSC) for supporting this research, Grant/Award Number: 202308080109. The Melville data has been provided by the city of Melville.

References

Abayowa, B. O., Yilmaz, A., Hardie, R. C., 2015. Automatic registration of optical aerial imagery to a LiDAR point cloud for generation of city models. *ISPRS Journal of Photogrammetry and Remote Sensing*, 106, 68–81.

Bodensteiner, C., Arens, M., 2012. Real-time 2D video/3D LiDAR registration. *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, IEEE, 2206–2209.

Borer, J., Tschirner, J., Ölsner, F., Milz, S., 2024. From chaos to calibration: A geometric mutual information approach to target-free camera LiDAR extrinsic calibration. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 8409–8418.

Bulatov, D., 2019. Alignment of building footprints using quasi-nadir aerial photography. *Scandinavian Conference on Image Analysis*, Springer, 361–373.

Bulatov, D., Burkard, E., Ilehag, R., Kottler, B., Helmholz, P., 2020. From multi-sensor aerial data to thermal and infrared simulation of semantic 3D models: Towards identification of urban heat islands. *Infrared Physics & Technology*, 105, 103233.

Council of City of Meville, 2017. Urban forest strategic plan 2017-2036: Plan a: City-controlled plan. Technical report, City of Melville, Perth, Australia.

Fan, H., Zipf, A., Fu, Q., Neis, P., 2014. Quality assessment for building footprints data on OpenStreetMap. *International Journal of Geographical Information Science*, 28(4), 700–719.

Fervers, F., Bullinger, S., Bodensteiner, C., Arens, M., Stiefelhagen, R., 2022. Continuous self-localization on aerial images using visual and LiDAR sensors. *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 7028–7035.

Guo, M., Wu, X., Qi, H., Zhang, Y., Chen, J., Wei, Y., Zhang, X., Shang, X., 2025. A methodology for sky-space-ground integrated remote sensing monitoring: A digital twin framework for multi-source Data-BIM integration in residential quality monitoring. *Journal of Building Engineering*, 102, 111976.

Haklay, M., 2010. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design*, 37(4), 682–703.

Han, S., Zhu, X., Wu, J., Cai, X., Yang, W., Yu, H., Xia, G.-S., 2025. DF-Calib: Targetless LiDAR-Camera Calibration via Depth Flow. *arXiv e-prints*, arXiv–2504.

Huang, R., Zheng, S., Hu, K., 2018. Registration of aerial optical images with LiDAR data using the closest point principle and collinearity equations. *Sensors*, 18(6), 1770.

Ilehag, R., Bulatov, D., Helmholz, P., Belton, D., 2018. Classification and representation of commonly used roofing material using multisensorial aerial data. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42, 217–224.

Leahy, J., Jabari, S., 2024. Enhancing Aerial Camera-LiDAR Registration through Combined LiDAR Feature Layers and Graph Neural Networks. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48, 25–31.

Lee, E., Kwon, Y., Kim, C., Choi, W., Sohn, H.-G., 2024. Multi-source point cloud registration for urban areas using a coarse-to-fine approach. *GIScience & Remote Sensing*, 61(1), 2341557.

Lv, F., Ren, K., 2015. Automatic registration of airborne LiDAR point cloud data and optical imagery depth map based on line and points features. *Infrared Physics & Technology*, 71, 457–463.

Shahraki, M., El-Rabbany, A., Elamin, A., 2024. UAV-based multi-sensor data fusion for 3D building detection. *Proceedings*, 110number 1, MDPI, 12.

Tamas, L., Kato, Z., 2013. Targetless calibration of a LiDAR-perspective camera pair. *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 668–675.

Wang, R., Xu, S., Dai, C., Xiang, J., Deng, Y., Tong, X., Yang, J., 2025a. MoGe: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 5261–5271.

Wang, R., Xu, S., Dong, Y., Deng, Y., Xiang, J., Lv, Z., Sun, G., Tong, X., Yang, J., 2025b. MoGe-2: Accurate monocular geometry with metric scale and sharp details.