3-D Point-Guided Aerial–Ground Image Matching for Robust Multiview Reconstruction

Yilin Xiao, Yu Yang, Siliang Du, Mingzhong Liu, Xu Chen, and Mingwei Sun

***Abstract*—Matching and aligning ground and aerial images are critical for enhancing the accuracy and completeness of 3-D reconstruction. However, significant differences in perspective and radiometric characteristics between aerial and ground images make this task highly challenging. Existing mesh-based approaches often overlook the geometric properties of 3-D points in the structure-from-motion model and suffer from limited track length. To address these issues, we propose a 3-D point-guided matching framework that leverages reconstructed 3-D points to guide the matching between aerial and ground images. Our method introduces a 3-D point-guided transformer to encode point coordinates into embeddings and integrate them into image features, enabling effective correspondence between synthetic aerial views and real ground images. In addition, we design a Transformer-based regression module to refine matching positions within local windows, improving the accuracy of aerial–ground correspondences. Our pipeline reduces matching errors, enables long-track correspondences, and facilitates robust multiview integration. Furthermore, we construct two challenging aerial–ground datasets to validate the effectiveness of our method in city-scale 3-D reconstruction. Extensive experiments on public benchmarks and our datasets demonstrate that our framework significantly outperforms state-of-the-art methods in both matching accuracy and reconstruction quality.**

***Index Terms*—3-D point-guided matching (PGM), 3-D reconstruction, aerial–ground image matching and alignment, transformer-based regression.**

I. Introduction

THE task of 3-D reconstruction plays a central role in remote sensing and urban modeling. In practice, 3-D reconstruction typically involves capturing both aerial and ground images, which provide complementary viewpoints. Aerial images offer a global view with clear rooftop structures, while ground images capture detailed façade information, especially in occluded or shadowed regions. Integrating these heterogeneous perspectives can significantly enhance the completeness and realism of the reconstructed scene. However, aligning aerial and ground images remains a challenging task due to large differences in perspective, scale, resolution, and illumination. This challenge motivates the problem of aerial–ground image matching, which aims to establish reliable correspondences between image features across views. Accurate matching is a prerequisite for effective multiview fusion and high-quality 3-D reconstruction [1], [2].

Existing approaches for aerial–ground matching can be roughly categorized into three types: direct matching, transform-based methods, and mesh-based pipelines. Direct matching [3] attempts to match features across perspectives directly, but suffers from severe viewpoint and radiometric differences. Transform-based methods [4], [5] rectify the views using DEM or ground planes, but are limited by low overlap and high computational cost.

In contrast, mesh-based methods [6], [7] employ structure-from-motion (SfM)-generated sparse models and use aerial meshes to render synthetic ground-view images. These synthetic views help bridge the viewpoint gap by enabling matching between ground images and synthetic views. However, existing mesh-based pipelines suffer from two main limitations. First, synthetic images often exhibit deformation and texture artifacts, which degrade matching reliability. Second, most methods ignore the geometric properties of 3-D points reconstructed in the SfM model. Consequently, they fail to exploit the long-term consistency of 3-D points across multiple views, leading to short track lengths and unstable matching. Track length, defined as the average number of images in which a given 3-D point is both detected and successfully matched, serves as a key metric for feature reliability in SfM. Longer tracks imply more robust matches and thus provide the foundation for high-precision, globally consistent 3-D reconstructions.

To overcome these limitations, we propose a novel mesh-based pipeline for 3-D point-guided aerial–ground matching, which explicitly incorporates the geometric information of SfM 3-D points to guide feature matching across views, as shown in Fig. 1. Our framework is composed of two key components. 1) A 3-D point-guided transformer (PGT) that encodes 3-D point coordinates into geometric embeddings and integrates them into image features, enabling synthetic-to-ground matching through both patch-level and point-level associations. 2) A Transformer-based regression module (TRM) that refines the initial correspondences by performing local regression within matching windows, thereby improving spatial precision between aerial and ground images. By leveraging the point-to-image visibility information and maintaining geometric consistency

Received 16 June 2024; revised 27 May 2025; accepted 23 September 2025. Date of publication 1 October 2025; date of current version 23 October 2025. *(Yilin Xiao and Yu Yang contributed equally to this work.) (Corresponding author: Siliang Du.)*

Yilin Xiao is with the The Hong Kong Polytechnic University, Hong Kong SAR 999077, China (e-mail: yilin.xiao@connect.polyu.hk).

Yu Yang is with the Institute of Cyber-Systems and Control, Zhejiang University, Hangzhou 310058, China (e-mail: yu.yang@zju.edu.cn).

Siliang Du, Mingzhong Liu, and Xu Chen are with Huawei Technologies Company, Ltd., Wuhan 430074, China (e-mail: dusi@whu.edu.cn; mingzhongliu@foxmail.com; chenxuflying@foxmail.com).

Mingwei Sun is with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430072, China (e-mail: mingweis@whu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2025.3616417

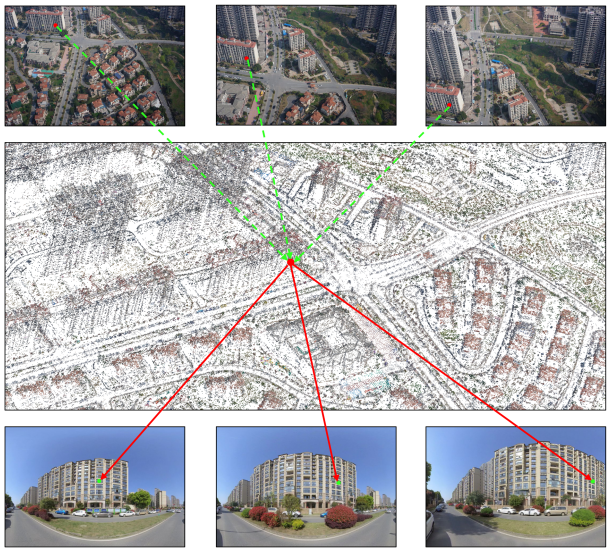


Fig. 1. *3-D point-guided aerial–ground image matching.* We match the 3-D points in the aerial model to the ground sequence, establishing a covisible relationship between aerial–ground sequences.

across views, our method facilitates the construction of longer and more reliable tracks, which are critical for high-fidelity 3-D reconstruction.

In practice, we first apply an SfM pipeline to the aerial image sequence to reconstruct 3-D points and generate a sparse mesh. Based on this, we render synthetic images from the ground perspective using the aerial mesh. These synthetic views serve as an intermediate domain for bridging the wide perspective gap between aerial and ground imagery. The PGT module takes as input the 3-D point coordinates and the synthetic-ground image pair, encoding the geometric cues into feature embeddings. It then performs hierarchical matching from patch-level to point-level under the supervision of 3-D geometry, allowing robust correspondence estimation despite appearance variation. Following this, the TRM module aggregates features within local windows around the matched points in both aerial and ground images. It uses a transformer attention mechanism to regress more accurate matching positions and eliminate local misalignments. By associating each 3-D point with multiple ground observations, our method naturally extends track lengths across views, leading to more consistent multiview correspondences and ultimately enhancing the completeness and accuracy of the final 3-D mesh reconstruction.

We evaluate our proposed methods on several benchmarks, i.e., Zeche, Centre [8], and SWJTU-BLD [6] datasets. In addition, we construct two more challenging datasets, namely, Huashan-BLD and Huashan-Low, using higher uncrewed aerial vehicles (UAVs) acquisition altitudes and ground panoramic imaging. These datasets primarily focus on high-rise and ground-level buildings, bringing us closer to the real-world conditions necessary for city-level 3-D mapping. The experiments demonstrate that our framework achieves more robust and accurate aerial–ground image-matching results, even in challenging scenarios with variations in perspective, resolution, and radiation. The core contributions of this article are summarized as follows.

1) We propose a novel mesh-based pipeline for 3-D point-guided aerial–ground image matching, which explicitly incorporates the geometric features of 3-D points reconstructed via SfM to guide and constrain the cross-view matching process.
2) We design a 3-D PGT for injecting 3-D geometry into image features to guide synthetic-to-ground matching, and a TRM for refining aerial–ground correspondences via local attention-based regression. This two-stage framework improves match accuracy and track consistency.
3) We construct two challenging real-world datasets, Huashan-BLD and Huashan-Low, featuring diverse viewpoints and scales. Extensive experiments show that our method outperforms existing approaches in both aerial–ground image matching and multiview 3-D mesh reconstruction.

II. Related Work

A. Local Feature Matching

Local feature matching is a crucial task widely employed in various remote sensing applications, including aerial–ground image alignment and fusion. Existing methods for local feature matching can be categorized into two main groups: detector-based and detector-free methods. In detector-based methods, the same feature detector and descriptor are applied to both images to facilitate the matching process. Among these methods, SIFT [9] has emerged as a highly influential detector, extensively utilized in computer vision and remote sensing tasks. RIFT [10] is a variant of sift, which is based on radiation-variation insensitive feature transform. HOPC [11] is a novel feature descriptor with the histogram of orientated phase congruency, which is based on the structural properties of images. LNIFT [12] presents a local normalization filter to convert original images into normalized images for feature detection and description. CFOG [13] is a pixelwise feature representation using orientated gradients of images, which is an extension of the pixelwise histogram of oriented gradient descriptor. SRIF [14] obtains the scales of keypoints by projecting them into a simple pyramid scale space, which largely reduces the complexity compared to traditional Gaussian scale space. ASLFeat [15] leverages the inherent feature hierarchy to enhance spatial resolution and capture low-level details, enabling accurate keypoint localization. D2Net introduces a single convolutional neural network that serves as both a dense feature descriptor and a feature detector. SuperPoint [16] presents a self-supervised training approach using homographic adaptation, which has gained significant popularity in recent times. Furthermore, SuperGlue [17] introduces a network with a flexible context aggregation mechanism based on attention, achieving remarkable performance by leveraging graph neural networks to learn feature matching. LightGlue [18] is an effective improvement over SuperGlue, especially in terms of memory and computation.

Detector-free methods offer a unified approach to describe and match local features without relying on a separate feature detector. Several notable methods have been proposed in this category. NCNet [19] introduces a unique approach to learning dense correspondences directly in an end-to-end manner. It utilizes 4-D cost volumes to enumerate all potential matches between images and applies 4-D convolutions to regularize the cost volume, promoting consensus among matches. GLU-Net [20] is based on the optical flow method, which can provide a dense matching effect in the case of significant transformation. LoFTR [21] describes the image by a transformer that first establishes pixelwise dense matches at a coarse level and later refines the matches. COTR [22] first downsamples each image with a CNN, and then the result is fed into a transformer along with the query. ASpanFormer [23] is built on a hierarchical attention structure, adopting a novel attention operation that is capable of adjusting attention span in a self-adaptive manner. DKM [24] proposes a kernel regression global matcher that warp refinement through stacked feature maps and depthwise convolution kernels.

Nonetheless, the significant disparity in perspective, radiation, and resolution between ground and aerial images poses a challenge, rendering both detector-based and detector-free methods incapable of ensuring a sufficient number of homologous points. Consequently, the track length between ground and aerial images is considerably limited. These issues motivate us to introduce a novel matching pattern to address these limitations.

B. Aerial–Ground Image Alignment and Fusion

Due to the significance of aerial–ground image alignment and fusion in various applications, such as 3-D reconstruction, virtual reality, and map construction, extensive research has been conducted in this area. Several notable studies have focused on transform-based and mesh-based methods. In the transform-based approach, Fanta-Jende et al. [5] presented an automatic coregistration method for mobile images and oblique aerial images. They utilize facade planes extracted from a sparse point cloud as projection surfaces to overcome the substantial perspective differences between the images. Wu et al. [4] proposed a novel integration method that combines automatic feature matching and bundle adjustment between ground and aerial images. Based on the integrated results, they further optimize the geometry and texture of 3-D models generated from aerial imagery. In the mesh-based approach, Gao et al. [7] employed sparse mesh-based image synthesis to match ground and aerial images. They filter putative point matches using geometrical consistency checks and geometrical model verification. Zhu et al. [6] proposed leveraging photogrammetric mesh models for aerial–ground image matching, which can be directly incorporated into off-the-shelf SfM and multiview stereo (MVS) solutions. The feature matching between synthetic aerial and ground images is conducted through descriptor searching and geometry-constrained outlier removal. Li et al. [3] proposed a dense correspondence learning-based SfM that is inspired by optical flow estimation.

Despite the advancements made by existing aerial–ground image alignment and fusion methods, challenges still persist in establishing accurate covisible relationships.

III. Method

A. Overview

The fusion of ground and aerial images plays a crucial role in enhancing the effectiveness of 3-D reconstruction. Therefore, it is essential to establish the covisible relationship between the two types of data through feature matching of ground and aerial images. In this article, we propose a mesh-based method that utilizes 3-D point-guided matching (PGM) and regression, as depicted in Fig. 2.

Our primary focus is on converting the aerial image into a synthetic color representation similar to the ground perspective. This transformation lays the groundwork for the subsequent task: establishing a robust matching relationship between the ground image and the synthesized color image, facilitated by the developed 3-D PGM module. After establishing this matching correlation, we project the relationship onto the aerial image domain. This projection enables us to delineate the regression region within the aerial context. Precise calculation of corresponding point coordinates is then achieved by applying the TRM, an integral component of our proposed methodology, ultimately resulting in accurate alignment.

According to our proposed workflow, the problem of aerial–ground image alignment and fusion can be defined. Given the sequence of ground image $\{I_{g_1}, \ldots, I_{g_n}\}$ and the sequence of aerial image $\{I_{a_1}, \ldots, I_{a_n}\}$, the tracks of the aerial sequence are matched to multiple ground images to obtain covisible matching points and the covisible relationship between the two image sequences. We can formalize the problem as follows:

$$\{p_{g_1}, \ldots, p_{g_n}\} = M_{a\to g}(\{p_{a_1}, \ldots, p_{a_n}\} \mid \{I_{g_1}, \ldots, I_{g_n}\}, \{I_{a_1}, \ldots, I_{a_n}\}) \quad (1)$$

where p_{a_n} is the nth 3-D point of the track in the aerial image and p_{g_n} is the corresponding feature point in the ground image.

B. Preprocessing

In our methodology, we adopt distinct processing pathways for ground and aerial images. Utilizing established SfM tools, such as Metashape [25], we derive camera poses and sparse point clouds for both image categories. To facilitate coarse alignment between the ground model and the aerial reference system, we utilize embedded GPS data from ground camera records, enabling coarse registration. Subsequently, we conduct meshing operations on the sparse aerial model, serving as a crucial intermediary step for subsequent image synthesis and matching. Synthesizing synthetic color images relies on the ground camera's pose, with rendering executed within the aerial mesh model. For efficiency and precision in image synthesis, we employ ray-tracing with Embree implementation.

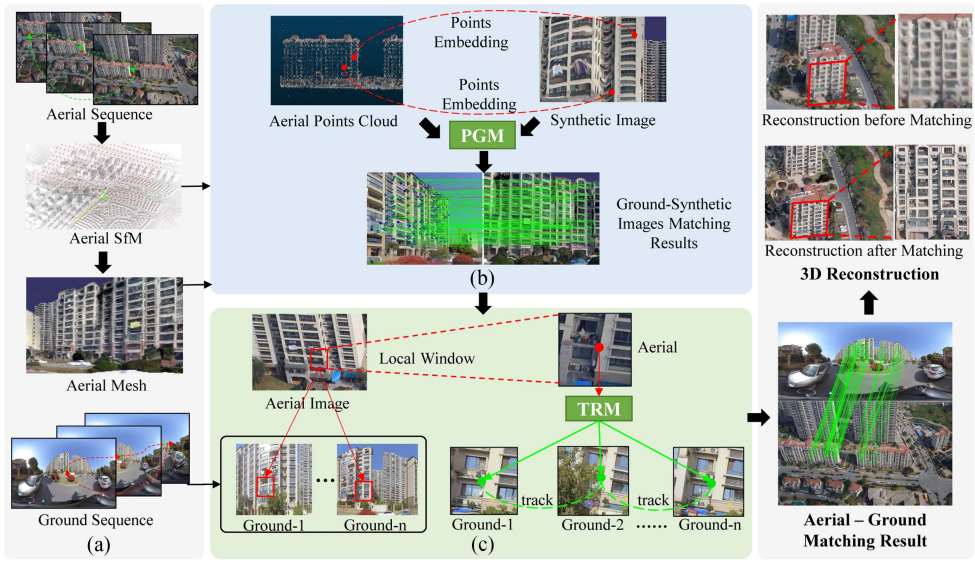


Fig. 2. *Workflow of the proposed method*. We first construct the SfM point cloud and mesh model from the aerial sequence and render synthetic images accordingly (Section III-B). Next, we perform 3-D PGM. by matching 3-D points in the synthetic color image to the ground image (Section III-C). We then determine the optimal regression region and refine the matched coordinates using a TRM (Section III-D). This process establishes accurate correspondences across aerial and ground views, forming reliable tracks for multiview fusion. As a result, our workflow significantly enhances the quality and completeness of 3-D reconstruction compared to using aerial data alone. (a) Input Preprocess. (b) 3D-Point-Guided Matching. (c) Refined Aerial–Ground Matching.

C. 3-D Point-Guided Matching

Upon completing the preprocessing phase, synthetic color images representing the aerial model from the ground perspective are generated. The subsequent aim is to establish correspondence between these synthetic color images and the ground images, facilitating the creation of a matching relationship between the 3-D points and the ground images. This article presents a novel matching approach centered on the shared utilization of 3-D points. The proposed methodology unfolds sequentially, starting with the extraction of image features using the 3-D PGT module. Subsequently, preliminary matches are established at the patch level, leading to the refinement of precise point-level correspondences.

1) Feature Extraction: We initially process the ground and synthetic images, denoted as I_g and I_s, through a feature extraction network to derive image features at both low-resolution and high-resolution levels. Specifically, we downsample each image using a CNN [26] to obtain a patch feature map, where each feature vector $\mathbf{f_p}$ corresponds to a subdivided patch $\mathbf{p}$ in either the patches of the ground image $\mathcal{P}^g$ or the patches of the synthetic image $\mathcal{P}^s$. Subsequently, we upsample the patch feature map to one-half of the original resolution using an FPN network [27] to derive the point feature $\mathbf{f_c}$ for each pixel $\mathbf{c}$. The feature extraction network acts as the image backbone, where the patch features $\mathbf{f_p}$ are further processed for patch matching, and $\mathbf{f_c}$ is utilized for point-to-point matching analysis.

2) 3-D PGT Module: In this work, a 3-D PGT module has been designed to take full advantage of the 3-D point coordinate information and reduce computation. First, we encode the patch coordinate with a 2-D extension of the standard positional encoding as follows:

$$\mathcal{PE}^m_{\mathbf{p}} = f(x,y)^m := \begin{cases} \sin(\omega_k \cdot x), & m = 4k \\ \cos(\omega_k \cdot x), & m = 4k+1 \\ \sin(\omega_k \cdot y), & m = 4k+2 \\ \cos(\omega_k \cdot y), & m = 4k+3 \end{cases} \tag{2}$$

where $\omega_k = \frac{1}{10000^{\frac{2k}{d}}}$ and (x,y) is the coordinate of a patch $\mathbf{p}$. d is the number of dimensions of the encoded feature, and $\mathcal{PE}^m_{\mathbf{p}}$ represents the value of its mth feature dimension. We add patch feature $\mathbf{f_p}$ with its positional encoding for patch $\mathbf{p}$, which enriches image features with spatial information. For each image, we stack all patch features as a matrix ${}^{(0)}\mathbf{x}$, representing the input to the PGT module.

Our PGT module consists of several attention layers. The ith attention layer can be described as mapping a query $\mathbf{q}_i$ and a set of key-value pairs $(\mathbf{k}_j, \mathbf{v}_j)$ to an output. $\mathbf{q}_i$, $\mathbf{k}_j$ and $\mathbf{v}_j$ can be expressed as follows:

$$\mathbf{q}_i = \mathbf{W}_1 {}^{(\ell)}\mathbf{x}_i + \mathbf{b}_1 \tag{3}$$

$$\begin{bmatrix} \mathbf{k}_j \\ \mathbf{v}_j \end{bmatrix} = \begin{bmatrix} \mathbf{W}_2 \\ \mathbf{W}_3 \end{bmatrix} {}^{(\ell)}\mathbf{x}_j + \begin{bmatrix} \mathbf{b}_2 \\ \mathbf{b}_3 \end{bmatrix} \tag{4}$$

where $\mathbf{W}_*$ and $\mathbf{b}_*$ are parameters to learn. ℓ stands for the ℓth attention layer. ${}^{(\ell)}\mathbf{x}$ represents the output from $(\ell-1)$th layer and input to ℓth layer.

Next, we obtain the message $\mathbf{m_i}$ by weighting and aggregation through the attention mechanism

$$\mathbf{m}_i = \sum_{j=1}^{|\mathcal{P}|} \alpha_{ij} \mathbf{v}_j \tag{5}$$

where the attention weight $\alpha_{ij} = \phi(\mathbf{q}_i)\phi(\mathbf{k}_j^\top)$ and $\phi(\cdot) = \mathrm{elu}(\cdot) + 1$. We employ linear attention [28], wherein the softmax operator is substituted with the product of two kernel functions, as a noteworthy approach. This technique significantly alleviates the memory and computational overhead associated with attention weights, particularly in scenarios where the image resolution is high. Finally, we output the feature for the next layer

$$ {}^{(\ell+1)}\mathbf{x}_i = {}^{(\ell)}\mathbf{x}_i + \mathrm{MLP}\left(\left[{}^{(\ell)}\mathbf{x}_i \| \mathbf{m}_{\mathcal{E}\to i}\right]\right) \tag{6} $$

where $[\cdot \| \cdot]$ denotes concatenation.

Based on the attention layer, we design a 3-D PGT module. First, we use the self-attention layer to process f_p^s and f_p^g, respectively, to obtain the information inside the features. Since we only care about the matching results of the patches where the 3-D points are located, we filter out the patches not containing 3-D points for f_p^s to reduce the useless computation. The filtered patch features are no longer computed with the global patch features for attention. Then, we add the 3-D positional encoding information of the 3-D points to them to generate the 3-D point-aware features. 3-D positional encoding can be described by the following equation:

$$ P_i^{\text{3-D}}[0] = (P_i^{\text{3-D}}[0] - x_{\min})/(x_{\max} - x_{\min}) \tag{7} $$

$$ P_i^{\text{3-D}}[1] = (P_i^{\text{3-D}}[1] - y_{\min})/(y_{\max} - y_{\min}) \tag{8} $$

$$ P_i^{\text{3-D}}[2] = (P_i^{\text{3-D}}[2] - z_{\min})/(z_{\max} - z_{\min}) \tag{9} $$

where $x_{\max}$, $y_{\max}$, and $z_{\max}$ are the maximum values of the *XYZ* coordinates and $x_{\min}$, $y_{\min}$, and $z_{\min}$ are the minimum values of the *XYZ* coordinates. Given a 3-D point $P^{\text{3-D}}$, we first normalize the *XYZ* coordinates separately. We then use the sine function to transform each dimensional coordinate value into a $\frac{C}{2}$-dimensional vector and then concatenate them together to obtain a $\frac{3C}{2}$-dimensional vector, where C is the number of channels. Then, the MLP consisted of two linear layers, and a ReLU activation reduces the vector dimension from $\frac{3C}{2}$ to C

$$ \begin{aligned} PE_i^{\text{3-D}} = \ & \mathrm{MLP}(\mathrm{Cat}(\mathrm{Sine}(P_i^{\text{3-D}}[0]), \mathrm{Sine}(P_i^{\text{3-D}}[1]) \\ & \mathrm{Sine}(P_i^{\text{3-D}}[2]))). \end{aligned} \tag{10} $$

We finally stack multiple self-attention layers and cross-attention layers to obtain richer interaction information between features.

3) Patch-to-Patch Matching: After the 3-D PGT module, we obtain the patch features for both the ground and synthetic images as $\{\mathbf{x}_i^g\}$ and $\{\mathbf{x}_j^s\}$. We assign a value of 0 to patch features for those filtered-out patches. We establish matches between ground and synthetic patches by solving an optimal transport problem following SuperGlue [17] using the Sinkhorn algorithm [29]. We compute pairwise patch similarity scores as

$$ \mathbf{S}_{i,j} = < \mathbf{x}_i^g, \mathbf{x}_j^s > \quad \forall (i,j) \in |\mathcal{P}^g| \times |\mathcal{P}^s| \tag{11} $$

where $< \cdot, \cdot >$ is the inner product. Then, we solve the optimal transport problem to obtain an assignment matrix $\mathbf{A}_{i,j}$ by maximizing the total score $\sum_{i,j} \mathbf{S}_{i,j}\mathbf{A}_{i,j}$. We use the same loss formulation for patch-to-patch matching as in [17].

4) Point-to-Point Matching Based 3-D Point-Guidance: With the establishment of the patch-to-patch assignment matrix $\mathbf{A}$, our focus shifts toward deriving point-to-point matches from it. The assignment matrix $\mathbf{A}$ functions as a repository for encapsulating correlations between patches in distinct images. For synthetic image patches accompanied by associated 3-D points, we strategically designate the ground image patch exhibiting maximum confidence as the reference domain for subsequent point-to-point matching.

For each pair of matching patches in the patch feature maps, we first locate their respective positions within the point feature maps $\mathbf{f}_\mathbf{c}^s$ and $\mathbf{f}_\mathbf{c}^g$. The synthetic image patches are positioned using the coordinates of their associated 3-D points, while the ground image patches are positioned based on the center coordinates of each patch. Subsequently, the application of ROIAlign [30] ensues, extracting subpixel windows of dimensions $w * w$ centered around the designated points. These cropped windows are then fed into the point-to-point matching module, which consists of an encoder and a decoder based on the transformer, resulting in the transformed point feature maps $\hat{\mathbf{f}}_\mathbf{c}^s$ and $\hat{\mathbf{f}}_\mathbf{c}^g$. Through a series of feature transformations, we calculate the correlation between the center vector of $\hat{\mathbf{f}}_\mathbf{c}^s$ and all vectors in $\hat{\mathbf{f}}_\mathbf{c}^g$. These correlation data are leveraged to generate a heatmap, quantifying the matching degree between each pixel in the ground image and the 3-D point in the synthetic image. Probability distribution expectations are computed, leading to the identification of ground image points corresponding to the 3-D points in the synthetic image. The culmination of this iterative process yields the comprehensive set of point-to-point correspondences denoted as $\mathcal{M}_c$. In the process of point-to-point matching, the identical loss formulation employed in [21] is adopted.

D. Refined Aerial–Ground Matching

1) Localization of Local Regression Region: After the 3-D PGM module, we establish the matching relationship between synthetic color images and ground images. Despite originating from a meticulously constructed aerial mesh model, the synthetic image still exhibits discrepancies when aligned with the ground image. These disparities arise from deformations, texture loss, and inherent noise within the mesh model, leading to deviations between the synthetic and actual ground images, as illustrated in Fig. 5. To address these discrepancies, we introduce an additional step aimed at extending the established matching relationship between the ground and synthetic images to the aerial images. This step not only extends the matching but also sets the stage for the precise regression of matching coordinates within the ground image, utilizing the regression region within the aerial image. A critical aspect of this process involves judiciously selecting the most suitable aerial image.

Given the aerial sparse point cloud and its corresponding mesh, each 3-D point can be unambiguously associated with a particular triangular face on the mesh. To determine the most suitable aerial image for further processing, we use the projected area of the face where the 3-D point resides on the aerial image as a selection criterion. When the aerial image's viewpoint is frontal and close to the 3-D mesh, the projection area tends to be

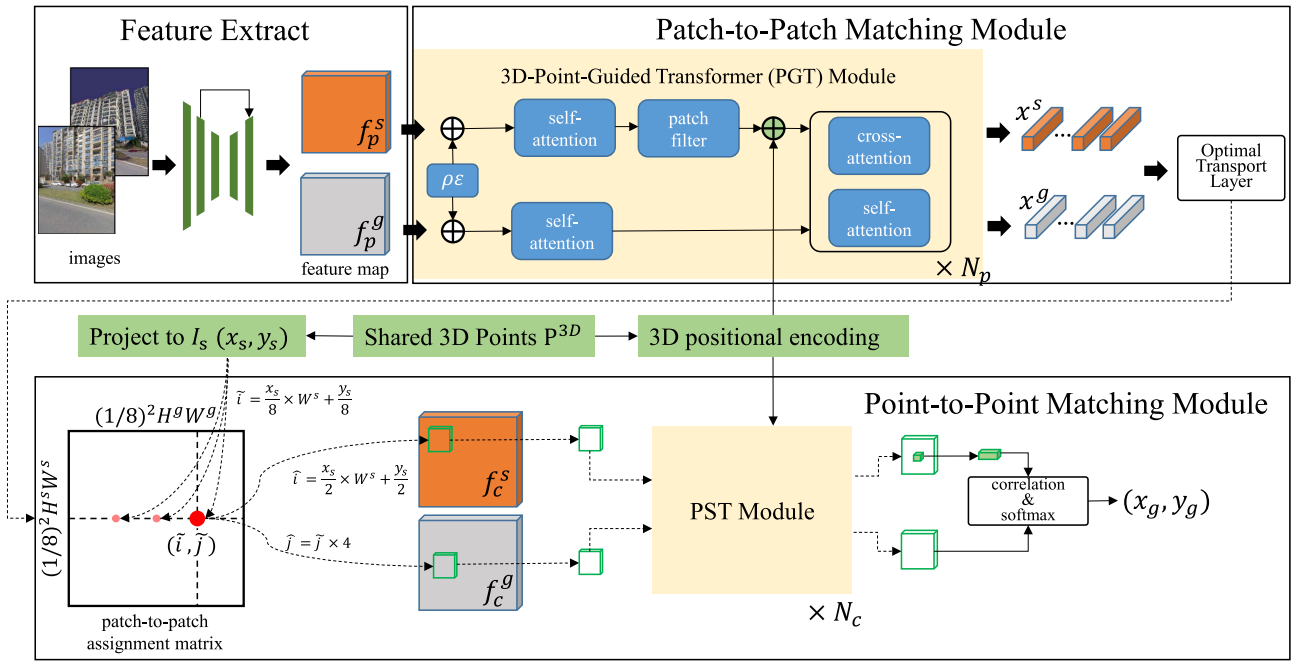
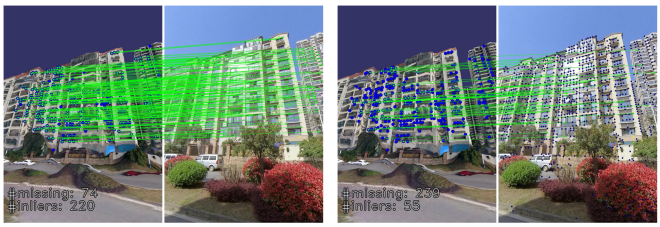


Fig. 3. *Proposed 3-D PGM method.* Our method is divided into feature extraction, patch-to-patch matching, and point-to-point matching. Before matching, we use CNN to extract the patch and point feature maps from the two types of images. Patch-to-patch matching involves leveraging the 3-D PGT to process the patch features within the images. Simultaneously, 3-D points are served as the regions of interest for patch filtering. The matching results are obtained through the optimal transport algorithm. In point-to-point matching, the shared 3-D point is further matched and regressed within the patch range to obtain the final matching result.



(a) (b)

Fig. 4. *Comparison of the matching effect between ground and synthetic images.* Our method has significantly more matching points than Superpoint + SuperGlue [16], [17]. For the Superpoint + SuperGlue method, we redescribe the image using Superpoint. (a) Ours. (b) Superpoint+SuperGlue.

Fig. 5. *Difference between synthetic and ground images.* These errors arise from deformations, texture loss, and noise in the mesh model, leading to discrepancies between the synthetic and actual images.

larger, indicating a more appropriate aerial image for subsequent operations. Formally, we define

$$S = \beta(\sigma, I_a) \tag{12}$$

where S represents the projection area, β represents the projection function, σ represents the face of 3-D points, and I_a represents the aerial image.

Once the optimal aerial image has been identified, we proceed to extract a window from it, referred to as the reprojection area. This window is cropped based on the coordinates on the aerial image corresponding to the specific 3-D point within the mesh model. We then reproject this window onto the ground image using depth and normal vector information. This two-step cropping and reprojection yields a tightly localized search region in the ground image, substantially improving the accuracy of downstream coordinate regression. The specific formula is as follows:

$$d_a = \frac{(t_0 - t_a)\overrightarrow{n}}{(R_a^{-1} m_a)\overrightarrow{n}} \tag{13}$$

$$m_g = R_g(R_a^{-1} d_a m_a + t_a - t_g) \tag{14}$$

where d_a is the depth information of the aerial image, m_a is the normalized coordinates of a point in the window of the aerial image, m_g is the normalized coordinates of the reprojected point in the ground image, $\overrightarrow{n}$ is the normal vector of the 3-D point, t_0 is the 3-D coordinate of the 3-D point, t_a and t_g are the camera positions of the aerial image and the ground image, respectively, R_a is the rotation matrix of the aerial image from world to the camera, and R_g is the rotation matrix of the ground image from world to the camera. To further guard against errors in the

ground image pose, we recentralize the reprojection window in the ground image by shifting its center to coincide with the original matching point.

2) Transformer-Based Regression in Windows: After propagating the matching relationship to the aerial image, we acquire both the aerial image window and its corresponding ground image window. The primary objective of the refinement stage is to achieve precise alignment between these windows. Currently, the refinement stage relies on template matching, which mainly utilizes pixel color information and overlooks the rich structural details available in higher dimensions. Consequently, this method fails to yield satisfactory results. To address this limitation, we introduce a regression technique based on transformer architecture, as shown in Fig. 3.

To derive comprehensive structural feature representations, we employ the lower layers of ResNet [26] as the backbone network. Specifically, we extract the feature map with 1024 channels after layer 3. To reduce computational complexity, we project the 1024-channel feature maps using 1×1 convolution to obtain 256-channel feature maps. These windows are then inputted into ResNet [26] to extract more informative features. Subsequently, we merge the feature maps from the two windows and integrate positional information through positional encoding. It is imperative to emphasize the significance of concatenating two feature maps. Rather than concatenating along the channel dimension, which might introduce artificial associations between features from the same pixel locations in each image, we concatenate along the spatial dimension. This spatial concatenation allows the features in each feature map to function as tokens, fostering global interactions and preserving spatial relationships between features from both images while fostering long-range dependencies.

The concatenated feature map, combined with 3-D point coordinates, serves as the input for the transformer model [31]. Our transformer architecture comprises two layers in both the encoder and decoder. Each encoder layer contains an eight-head self-attention module, while each decoder layer incorporates an eight-head encoder–decoder attention module. This configuration effectively integrates local and global information through self-attention and cross-attention mechanisms, resulting in a 256-channel feature map as the transformer's output.

We utilize max pooling to extract a global latent vector representing the corresponding point's location. To regress the coordinates of this corresponding point, we employ a three-layer MLP. Specifically, the input to the coordinate regressor is a 768-dimensional vector obtained by concatenating the two global latent vectors from the input windows and the positional-encoded point. This refined regression process leverages the transformer's capacity to integrate both structural features and positional information, thereby enhancing the precision of the refinement stage.

We use the l_2 loss for measuring the coordinate regression error

$$\mathcal{L} = \frac{1}{N} \sum_{i} \left\| c_{gt}^{i} - c_{r}^{i} \right\|_{2}^{2} \tag{15}$$

where N is the number of training samples, c_{gt}^{i} represents the real coordinates of the ith sample, and c_{r}^{i} represents the coordinates of the ith sample output by the network.

After regression, precise matching relationships between aerial and ground images are established. This process enables 3-D points in aerial images to be matched and observed across multiple distinct ground images, thereby facilitating the construction of visual tracks and increasing track length.

E. Implementation Details

Our models are implemented using the PyTorch framework, and the training of the 3-D PGM model is conducted on the MegaDepth dataset [32]. Training employs the Adam optimizer with an initial learning rate of 1×10^{-4} and a batch size of 64. Convergence is typically achieved after approximately 20 h of training, distributed across eight GeForce RTX 3090Ti GPUs. The parameter N_p for the PGT module in patch-to-patch matching is set to 4, while N_c for the PGT module in point-to-point matching is set to 1. In addition, we train the transformer-based regression model using data from the MegaDepth dataset. This model is optimized using the Adam optimizer with an initial learning rate of 1×10^{-5} and a batch size of 16 over a course of 300 k iterations.

IV. Experiment

A. Dataset Description

As shown in Table I, the evaluation of this study encompasses five distinct scenes, consisting of two easy scenes (Centre and Zeche), one normal scene (SWJTU-BLD), and two challenging scenes (Huashan-BLD and Huashan-Low). The Centre and Zeche datasets [8] were acquired by ISPRS in Dortmund and Zurich, respectively. These datasets include aerial images captured around specific buildings at varying heights, ranging from 18 to 73 m and 7 to 25 m above ground level, respectively. The SWJTU-BLD dataset [6] was collected at the Southwest Jiaotong University, with aerial images obtained at an approximate altitude of 140 feet above ground level.

However, the development of city-scale 3-D maps necessitates an augmented efficiency in map data collection. Elevating the operational altitude of UAVs and complementing them with panoramic ground imaging emerges as a highly effective means to substantially enhance the data collection efficiency. Therefore, as shown in Fig. 6, we constructed two additional datasets, namely, Huashan-BLD and Huashan-Low. The Huashan-BLD dataset comprises 1815 aerial images acquired using the DJI Matrice 300 RTK, while 451 corresponding ground images were captured using the vehicle-mounted Insta360 Pro2 panoramic camera. The dataset primarily focuses on high-rise buildings. In contrast, the Huashan-Low dataset includes aerial images that are consistent with the Huashan-BLD dataset. However, it features 560 ground images collected using the handheld Insta360 OneR panoramic camera, specifically focusing on ground-level buildings. Notably, both datasets surpass their predecessors in terms of scale and coverage, with aerial collection conducted at an altitude of nearly 300 m. Consequently, these datasets

TABLE I
DETAILED DESCRIPTIONS OF THE FIVE DATASETS USED FOR EVALUATIONS

Dataset	Aerial Altitude (m)	Images		Resolution	
		Aerial	Ground	Aerial	Ground
Zeche [6]	7–25	172	147	6000 × 4000	6000 × 4000
Centre [6]	18–73	146	204	6000 × 4000	6000 × 4000
SWJTU-BLD [4]	140	207	88	6000 × 4000	6000 × 4000
Huashan-BLD	270	1815	451	6000 × 4000	6080 × 3040
Huashan-Low			560		6080 × 3040

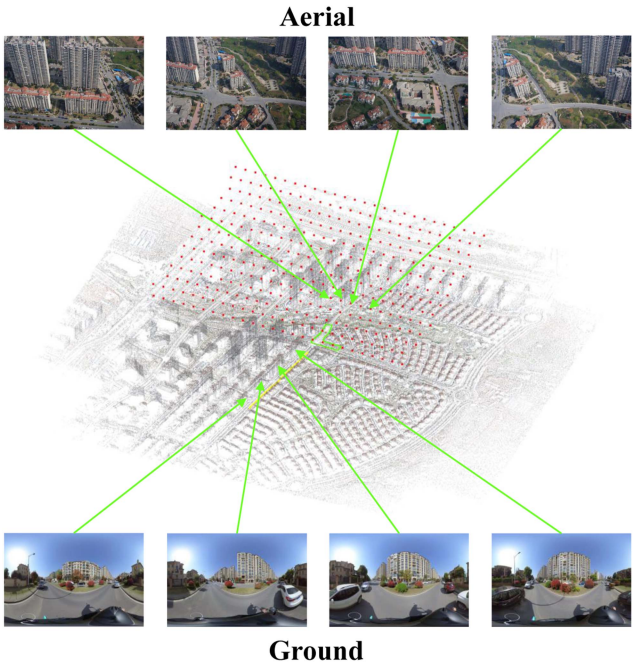


Fig. 6. *Huashan dataset.* The dataset features aerial imagery captured at altitudes of approximately 300 m, significantly higher than those in existing datasets. Ground-level images were collected using a panoramic camera to provide wide field-of-view coverage.

exhibit more pronounced variations in perspective, radiation, and overall complexity, providing a more rigorous evaluation environment.

The experiment conducted in this article involves the following preprocessing steps. First, the ground image and GPS data are employed to construct the SfM project, resulting in the acquisition of initial poses. Second, the aerial images are utilized to build SfM and MVS projects, thereby obtaining poses and meshes. It is pertinent to highlight that not all methods under evaluation have undergone training on these specific datasets, ensuring equitable and unbiased comparisons.

B. Evaluation of Reconstruction

We validate the effectiveness of our approach through reconstruction experiments involving the incorporation of aerial–ground correspondences into MetaShape [25]. This integration results in the refinement of ground image poses and the generation of an aerial–ground texture model. Our evaluation commences with examining the accuracy of aligned ground image poses, which is conducted through checkpoint error analysis.

Our primary objective revolves around the alignment of ground images with their aerial counterparts within the MetaShape project. Consequently, postfusion, the optimization process is exclusively dedicated to refining the ground image poses. While the Centre and Zeche datasets readily employ provided checkpoints, the SWJTU-BLD, Huashan-BLD, and Huashan-Low datasets necessitate a meticulous manual selection of 4–5 observation points on the aerial texture model. These points are chosen to ensure equitable visibility from the ground, effectively serving as checkpoints for evaluation. Our method consistently demonstrates superior performance across all datasets, as corroborated by the comparative analysis presented in the accompanying Table II. Furthermore, as the altitude of the aerial images increases, the advantages of our approach become progressively more pronounced. Particularly noteworthy are the remarkably low checkpoint errors achieved by our method on the Huashan-BLD and Huashan-Low datasets, where alternative methods fail to match. Fig. 7 demonstrates our matching advantages in various scenarios.

Beyond the quantitative evaluation based on checkpoint errors, we also qualitatively compare the effects of texture models constructed differently. As a reference comparison, we also explore the outcome of utilizing exclusively aerial data for texture model generation. Fig. 9 shows the comparison results. The first and second rows are the texture models generated by aerial data and fusion data, respectively. Selected segments of the models are magnified to highlight discrepancies more prominently. It is discernible from the comparison that the texture model derived from the fusion of data exhibits more comprehensive details. This enhanced detail manifests in the model's ability to distinctly depict facets, such as building facades, roof structures, and ground surfaces. In contrast, the texture model solely based on aerial data lacks ground information and appears less detailed, particularly in finer aspects.

C. Evaluation of Ground and Synthetic Images Matching

In this section, we compare several feature descriptions and matching methods on ground images and synthetic images. We also evaluate the performance of our proposed coarse matching strategy. Our task is to match 3-D points on the aerial model to the ground image. We already have preexisting feature points for this task. We only need to redescribe them and match them to the ground image. We choose SIFT + NN [9], D2Net + NN [35],

TABLE II
COMPARISON OF CHECKPOINT ERRORS ON THE FIVE DATASETS

Dataset	Error (cm)	Centre	Zeche	SWJTU-BLD	HuaShan-BLD	HuaShan-Low
RenderMatch* [4]	X	3.0	1.2	3.8	-	-
	Y	2.2	2.1	9.7	-	-
	Z	2.0	2.0	12.0	-	-
	Checkpoint error	4.3	3.1	15.9	-	-
LDCF* [1]	X	3.5	1.5	4.5	-	-
	Y	2.7	2.4	9.3	-	-
	Z	2.9	3.0	14.0	-	-
	Checkpoint error	5.3	4.1	17.4	-	-
Metashape [23]	X	8.4	2.3	-	-	-
	Y	5.5	1.7	-	-	-
	Z	4.9	1.1	-	-	-
	Checkpoint error	11.2	3.1	-	-	-
ColMap [31]	X	8.8	3.2	-	-	-
	Y	6.4	1.8	-	-	-
	Z	5.4	1.3	-	-	-
	Checkpoint error	11.7	3.8	-	-	-
PixSfM [32]	X	7.6	1.8	-	-	-
	Y	5.0	0.9	-	-	-
	Z	4.8	0.3	-	-	-
	Checkpoint error	10.3	2.1	-	-	-
Ours	X	2.3	0.4	2.6	6.8	7.5
	Y	0.7	1.1	3.2	6.2	6.1
	Z	2.2	0.9	2.0	2.7	4.7
	Checkpoint error	**3.3**	**1.5**	**4.6**	**9.6**	**13.4**

3-D coordinates of checkpoints for centre and zeche are provided with the datasets. The 3-D coordinates of the checkpoints for the remaining three datasets were taken from manually selected locations on the aerial texture model. The symbol "-" indicates missing data due to a failed match, and the symbol "*" indicates that the code of the method cannot be obtained or run and is re-implemented by us.

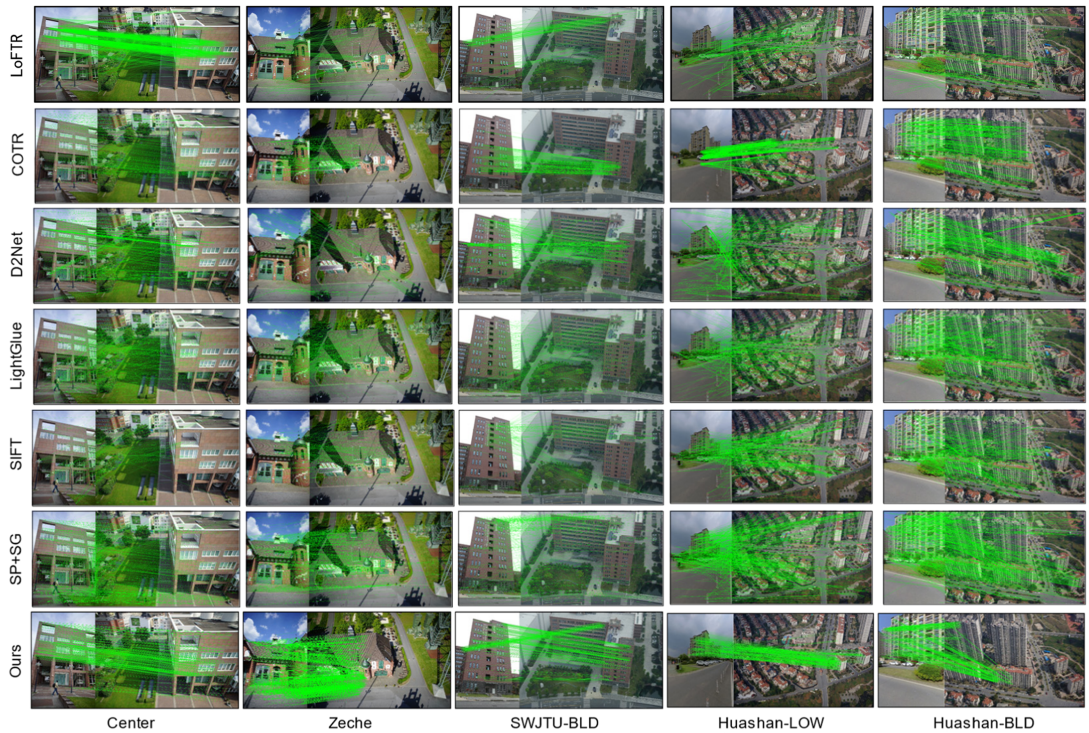


Fig. 7. *Aerial–ground image matching results.* Our method produces more accurate matches across various scenarios, outperforming both direct matching and mesh-based methods. These results highlight the effectiveness of our approach in bridging the modality gap between aerial and ground images.

TABLE III
COMPARISON OF MATCHING RESULTS BETWEEN GROUND AND SYNTHETIC IMAGES

Method	Easy		Normal	Hard	
	Zeche	Centre	SWJTU-BLD	Huashan-BLD	Huashan-Low
RenderMatch* [4]	36.1	31.1	25.0	19.8	10.9
SIFT+NN [7]	23.0	26.0	16.3	-	-
D2Net+NN [33]	217.4	207.7	104.0	156.0	88.7
ASLFeat+NN [13]	141.4	138.2	131.9	124.8	111.6
SP+SG [14], [15]	321.8	308.1	207.3	257.6	190.7
Lightglue [16]	341.2	321.1	218.7	270.6	200.3
LoFTR [19]	576.3	197.7	135.2	83.2	62.5
ASpanFormer [21]	565.5	192.4	127.8	79.9	60.1
Ours	**2108.3**	**1540.0**	**1449.0**	**1524.4**	**1189.4**

We use the number of matched interior points after PnP as evaluation metrics. The symbol "-" indicates missing data due to a failed match.

Fig. 8. *Qualitative example of regression results on the Huashan dataset*. Our method achieves the highest accuracy, producing predictions that are consistently closest to the ground truth compared to other methods.

TABLE IV
COMPARISON OF REGRESSION RESULTS

Method	errors (px)	time cost (ms)
NCC	8.71	2.1
PWC-Net [34]	5.53	87.2
GLU-Net [18]	2.82	87.4
FlowFormer [35]	2.57	90.1
COTR [20]	2.28	80.1
Ours	**1.27**	**41.2**

To demonstrate the effectiveness of the proposed TRM module.

ASLFeat + NN [15], superpoint + superglue [16], [17], and RenderMatch [6] for this purpose. While COTR [22] represents another proficient matching approach capable of aligning designated feature points, we have chosen not to incorporate it in this article due to its comparatively slower processing speed. We only match each ground image to its corresponding synthetic image. We use the number of matched interior points after PnP as evaluation metrics. For the detector-based methods, we extract 4096 feature points on the synthetic image. For ours and superpoint + superglue methods, which have a limit on the image size, we limit the maximum edge of the image to 1600. By the way, we still included two state-of-the-art (SOTA) detector-free methods [21], [23] for comparison. Table III and Fig. 4 shows the comparison results. The proposed method achieved the most matching points on five datasets. Especially as the difficulty of the scenario increases, the advantages of the proposed method become more prominent.

D. Evaluation of Ground and Aerial Images Regression

In order to verify the advantages of the TRM proposed in this article, we randomly select 1000 air–ground image matching pairs from the above datasets and use the matching pair results as the center to crop a 64 × 64 area window on the aerial image and the ground image, respectively. We manually labeled the matching results for each pair of cropping windows. In the task of this article, we only need to match the coordinates of the center point of the aerial cropping window on the ground cropping window. Therefore, we compare our method with the least squares and SOTA optical flow methods. We use the Euclidean distance between the matching result and the ground truth to evaluate the accuracy and use the matching time to evaluate the speed. For the COTR method, since the image size is below 256 × 256, we only iterate once.

As demonstrated in Table IV, the experimental results indicate that our algorithm outperforms the SOTA optical flow method in terms of accuracy in this scenario while achieving a twofold speed increase. Fig. 8 displays the matching result image obtained using the centralized method. Upon inspection of the matching result image, it becomes evident that our approach significantly improves accuracy, particularly in weakly textured and blurred areas, compared to the traditional NCC method.

E. Evaluation of Track Length

In this section, we compare the track length between our proposed method and SOTA techniques. As presented in Table V, our method consistently achieves a significantly greater track length than the other methods across all datasets. Even on simpler datasets, such as Zeche, where RenderMatch exhibits a

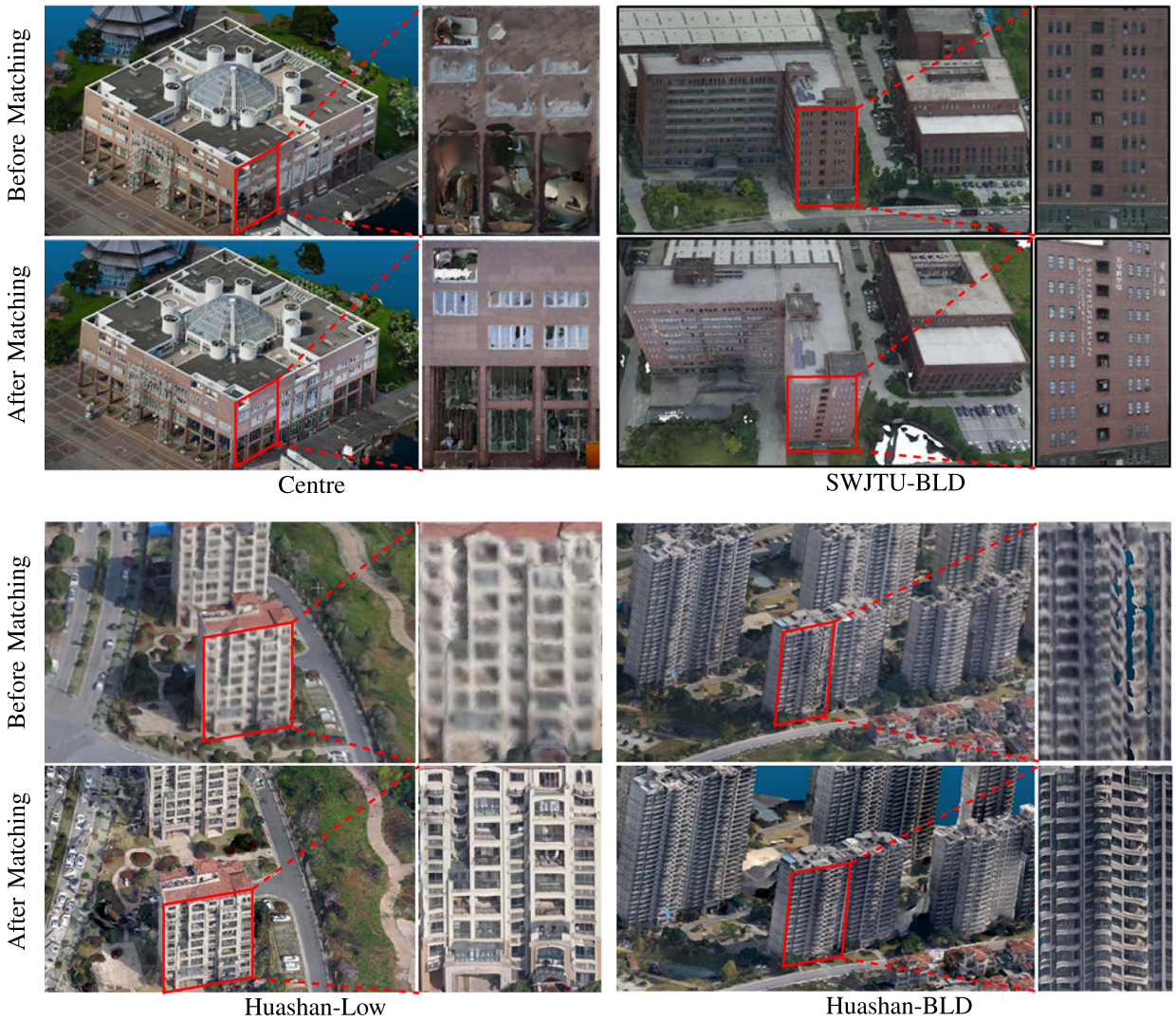


Fig. 9. *Comparison of the textured mesh models generated from only aerial and aerial–ground images.* The right column of each subfigure is an enlargement of the regions highlighted by the rectangles.

TABLE V
COMPARISON OF AVERAGE TRACK LENGTH

Method	Zeche	Centre	SWJTU-BLD	Huashan-BLD	Huashan-Low
ColMap [31]	3.23	2.52	2.37	2.08	2.12
PixSfM [32]	3.41	2.57	2.43	2.11	2.14
LDCF [1]	4.12	3.00	2.78	2.15	2.10
RenderMatch [4]	4.31	2.92	2.99	2.13	2.20
Ours	**10.83**	**8.82**	**9.15**	**7.98**	**7.96**

TABLE VI
COMPARISON OF TIME COST

Method	Rendering	Pairwise Match	Refinement
RenderMatch [4]	3.8 s	4.3 s	5.3 s
Ours	**1.7 s**	**0.1 s**	**2.4 s**

Our method achieves significantly lower time consumption than the SOTA Approach while achieving substantially higher accuracy.

decent track length, our proposed method outperforms it by a considerable margin. In challenging datasets, such as Huashan, most methods fail to match an adequate number of points, resulting in a limited track length. Conversely, our proposed method demonstrates robust performance in these complex scenarios. We attribute this success to two key factors. First, our method fully capitalizes on the available 3-D point information within the existing SfM model, enabling matching based on known 3-D points. Second, we design a TRM to establish accurate and more covisible relationships, which determines accurate coordinates through global optimization.

F. Evaluation of Time Cost

In this section, we discuss the efficiency of the proposed method and some tricks we used when implementing the algorithm. The deep learning-based matching algorithm can perform inference on the GPU, significantly improving the execution efficiency of the aerial–ground matching algorithm proposed in this article. The size of the texture model mainly affects the time of this step. However, the rendering speed can be accelerated by chunking and layering the model. In the coarse match phase, we only match the synthetic image with one ground image, so the time complexity of this step is $O(n)$. At the same time, we

TABLE VII
ABLATION STUDY

Ablation module				Checkpoint Error (cm)				
Baseline	PGT	TRM	LRR	Centre	Zeche	SWJTU-BLD	Huashan-BLD	Huashan-Low
✓				8.55	7.71	11.87	16.79	19.70
✓	✓			6.62	5.93	8.48	12.41	14.37
✓	✓	✓		3.57	1.75	4.79	3.90	4.69
✓	✓	✓	✓	**3.3**	**1.5**	**4.58**	**3.63**	**4.42**

Evaluated modules are 3-D PGT, TRM, and localization of regression region (LRR). We evaluate the effect by comparing checkpoint errors on different datasets.

limit the size of the longest side of the image to 1600. Although the execution speed of the coarse match phase correlates with the quantity of matched feature points, the computational load associated with this step remains relatively modest. In addition, given that all computations can be efficiently executed on the GPU through matrix operations, the time taken for this phase constitutes the shortest duration within the entire process. The count of matched feature points similarly influences the pace at which the refinement phase unfolds. Nonetheless, owing to our streamlined design and the compact dimensions of the cropped area image, we can enhance efficiency through concurrent execution on the GPU.

In this article, we measure the efficiency of the algorithm using quantitative metrics. We randomly select two images from each of the five datasets for aerial–ground matching and then record the time consumption of each step. As shown in Table VI, our method significantly outperforms the SOTA method in time cost at each step, especially in the pairwise match step. It is worth noting that our proposed matching pattern, which directly matches 3-D points from the aerial model with the ground image, offers a distinct advantage compared to SOTA methods. Unlike traditional approaches that require the aerial image and pose as input for subsequent pose optimization, our method eliminates the need for this step. As a result, the amount of data involved is significantly reduced, leading to a substantial improvement in efficiency.

G. Ablation Study

To comprehensively evaluate the effectiveness of each module in our proposed framework, we conduct ablation experiments on multiple datasets, including Centre, Zeche, SWJTU-BLD, Huashan-BLD, and Huashan-Low. The baseline configuration solely comprises the matching components of our framework, replacing the 3-D PGT with a conventional transformer module. The Localization of the regression region module represents our proposed method for selecting appropriate aerial images for regression. As shown in Table VII, incorporating each additional proposed module beyond the baseline consistently reduces checkpoint errors across all datasets. These results strongly support the effectiveness and validity of each module within our framework.

V. DISCUSSION

Potential impact: The proposed 3-D PGM pipeline represents a significant advancement in aerial–ground image matching for 3-D reconstruction, and performs particularly well in urban and complex environments. By integrating geometric features from SfM point clouds into the synthetic image domain and introducing a two-stage transformer-based matching and regression strategy, the method effectively bridges the modality gap between aerial and ground views. This integration enables the generation of longer and more reliable feature tracks, which are essential for constructing high-fidelity 3-D models. The improved robustness and accuracy of image matching in the presence of perspective, resolution, and radiometric differences make the approach highly applicable to real-world, large-scale scenarios, such as city-scale digital twins, urban planning, and autonomous navigation. Moreover, the introduction of the Huashan-BLD and Huashan-Low datasets offers valuable benchmarks for future research and practical deployment under challenging conditions.

Known limitations: Despite its strengths, several limitations should be acknowledged. First, the method is inherently dependent on the quality and completeness of the SfM-generated point cloud and mesh model. Inaccuracies in this initial reconstruction phase can propagate through the pipeline, leading to degraded matching performance due to sparse or noisy 3-D geometry. Second, although the proposed approach enhances matching across sequences, it still relies on manually designed pipeline components and heuristic settings (e.g., search window selection), which may constrain its generalizability to other domains or challenging environments, such as rural areas or textureless surfaces. Lastly, the framework assumes adequate overlap between aerial and ground views, an assumption that may not hold in sparsely captured or unevenly distributed datasets. To overcome these limitations, future work will explore the integration of 3-D Gaussian splatting [38] to enable an end-to-end aerial–ground matching pipeline, which is expected to significantly improve model rendering quality and robustness.

VI. CONCLUSION

In this article, we introduce a robust framework for precise alignment and fusion of aerial and ground images, addressing challenges, such as perspective disparities, radiation variations, and limited track length. Our approach presents a novel matching pattern that utilizes mesh rendering to align 3-D points from the aerial model with the ground image. Specifically, we propose the 3-D PGM method, which initially extracts features through the 3-D PGT, integrating geometric features of 3-D points into image space. In addition, we introduce a TRM to establish accurate corresponding relationships, enabling precise determination of matching point coordinates through

local optimization. Experimental results across multiple datasets demonstrate the significant superiority of our framework over SOTA methods, resulting in enhanced reconstruction quality and improved accuracy. We foresee our method playing a pivotal role in remote sensing, environmental monitoring, and various other scenarios.

REFERENCES

[1] Y. Xiao, S. Du, X. Chen, M. Liu, and M. Sun, "Dualattention-transformer-based semantic reranking for large-scale image localization," *Appl. Intell.*, vol. 54, no. 9-10, pp. 6946–6958, May 2024, doi: 10.1007/s10489-024-05539-2.

[2] S. Du, Y. Xiao, J. Huang, M. Sun, and H. Liu, "Guided local feature matching with transformer," *Remote Sens.*, vol. 15, no. 16, 2023, Art. no. 3989, doi: 10.3390/rs15163989.

[3] H. Li, A. Liu, X. Xie, H. Guo, H. Xiong, and X. Zheng, "Learning dense consistent features for aerial-to-ground structure-from-motion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 5089–5102, 2023.

[4] B. Wu, L. Xie, H. Hu, Q. Zhu, and E. Yau, "Integration of aerial oblique imagery and terrestrial imagery for optimized 3 D modeling in urban areas," *ISPRS J. Photogrammetry Remote Sens.*, vol. 139, pp. 119–132, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271618300613

[5] P. Fanta-Jende, F. Nex, G. Vosselman, and M. Gerke, "Co-registration of panoramic mobile mapping images and oblique aerial images," *Photogrammetric Rec.*, vol. 34, no. 166, pp. 148–173, 2019. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/phor.12276

[6] Q. Zhu, Z. Wang, H. Hu, L. Xie, X. Ge, and Y. Zhang, "Leveraging photogrammetric mesh models for aerial–ground feature point matching toward integrated 3 D reconstruction," *ISPRS J. Photogrammetry Remote Sens.*, vol. 166, pp. 26–40, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271620301507

[7] X. Gao, S. Shen, Y. Zhou, H. Cui, L. Zhu, and Z. Hu, "Ancient chinese architecture 3D preservation by merging ground and aerial point clouds," *ISPRS J. Photogrammetry Remote Sens.*, vol. 143, pp. 72–84, 2018.

[8] F. Nex, M. Gerke, F. Remondino, H. Przybilla, M. Bäumker, and A. Zurhorst, "ISPRS benchmark for multi-platform photogrammetry," *ISPRS Ann. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. 2, no. 3, 2015, Art. no. 135.

[9] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, pp. 91–110, 2004.

[10] J. Li, Q. Hu, and M. Ai, "RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform," *IEEE Trans. Image Process.*, vol. 29, pp. 3296–3310, 2020.

[11] Y. Ye, J. Shan, L. Bruzzone, and L. Shen, "Robust registration of multimodal remote sensing images based on structural similarity," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2941–2958, May 2017.

[12] J. Li, W. Xu, P. Shi, Y. Zhang, and Q. Hu, "LNIFT: Locally normalized image for rotation invariant multimodal feature matching," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5621314.

[13] Y. Ye, L. Bruzzone, J. Shan, F. Bovolo, and Q. Zhu, "Fast and robust matching for multimodal remote sensing image registration," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9059–9070, Nov. 2019.

[14] J. Li, Q. Hu, and Y. Zhang, "Multimodal image matching: A scale-invariant algorithm and an open dataset," *ISPRS J. Photogrammetry Remote Sens.*, vol. 204, pp. 77–88, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271623002277

[15] Z. Luo et al., "Aslfeat: Learning local features of accurate shape and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, Jun. 2020, pp. 6588–6597.

[16] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Salt Lake City, UT, USA, Jun. 2018, pp. 337–33712.

[17] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, Jun. 2020, pp. 4937–4946.

[18] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys, "LightGlue: Local feature matching at light speed," in *Proc. IEEE/CVF Inter. Conf. Comput. Vis.*, 2023, pp. 17627–17638.

[19] I. Rocco, M. Cimpoi, R. Arandjelović, A. Torii, T. Pajdla, and J. Sivic, "Ncnet: Neighbourhood consensus networks for estimating image correspondences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 2, pp. 1020–1034, Feb. 2022.

[20] P. Truong, M. Danelljan, and R. Timofte, "GLU-Net: Global-local universal network for dense flow and correspondences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, 2020, pp. 6257–6267.

[21] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "Loftr: Detector-free local feature matching with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Nashville, TN, USA, Jun. 2021, pp. 8918–8927.

[22] W. Jiang, E. Trulls, J. Hosang, A. Tagliasacchi, and K. M. Yi, "Cotr: Correspondence transformer for matching across images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Montreal, QC, Canada, Oct. 2021, pp. 6187–6197.

[23] H. Chen et al., "Aspanformer: Detector-free image matching with adaptive span transformer," in *Proc. Computer Vision–ECCV*, 2022. pp. 20–36.

[24] J. Edstedt, I. Athanasiadis, M. Wadenbäck, and M. Felsberg, "DKM: Dense kernelized feature matching for geometry estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Vancouver, BC, Canada, Jun. 2023, pp. 17765–17775.

[25] Agisoft, "Agisoft metashape," 2019. [Online]. Available: https://www.agisoft.com/

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, 2016, pp. 770–778.

[27] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 936–944.

[28] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, "Transformers are RNNs: Fast autoregressive transformers with linear attention," in *Proc. 37th Inter. Conf. Mach. Learn.*, H. D. III and A. Singh, Eds., vol. 119, Jul. 2020, pp. 5156–5165.

[29] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *Proc. Adv. Neural Infor. Proces. Syst.*, vol. 26, 2013, pp. 2292–2300.

[30] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Inter. Conf. Computer Vis.*, 2017, pp. 2961–2969.

[31] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. 9th Int. Conf. Learn. Representations*, May 2021, pp. 3–7. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[32] Z. Li and N. Snavely, "Megadepth: Learning single-view depth prediction from internet photos," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 2041–2050.

[33] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 4104–4113.

[34] P. Lindenberger, P.-E. Sarlin, V. Larsson, and M. Pollefeys, "Pixel-perfect structure-from-motion with featuremetric refinement," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Montreal, QC, Canada, 2021, pp. 5967–5977.

[35] M. Dusmanu et al., "D2-Net: A trainable CNN for joint description and detection of local features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, Jun. 2019, pp. 8084–8093.

[36] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 8934–8943.

[37] Z. Huang et al., "Flowformer: A transformer architecture for optical flow," in *Proc. Computer Vision–ECCV*, 2022, pp. 668–685.

[38] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3D Gaussian splatting for real-time radiance field rendering," *ACM Trans. Graph.*, vol. 42, no. 4, pp. 1–14, 2023.