Image Anal Stereol 2026;45:123-140 doi: 10.5566/ias.3932
Original Research Paper

GEOMETRY-AWARE FEATURE ENHANCEMENT WITH LINEAR ATTENTION FOR ROBUST UAV PHOTOGRAMMETRIC RECONSTRUCTION UNDER WEAK AND ISOMORPHIC TEXTURES

XIAOCONG JIANG[1], TIANYANG LUO[2],✉
[1]School of Business, Suzhou Industrial Park Institute of Vocational Technology, Suzhou 215123, China.
[2]Hezhou University, Hezhou 542899, Guangxi, China.
e-mail: roy@ivt.edu.cn, luotianyang@hzxy.edu.cn
(Received February 23, 2026; revised June 13, 2026; accepted June 14, 2026)

ABSTRACT

UAV photogrammetry constitutes a fundamental technology for the lifecycle maintenance and digital twin construction of large-scale transport infrastructure. However, standard Structure-from-Motion (SfM) pipelines frequently falter in scenarios characterized by weak textures, such as asphalt, and repetitive patterns. This deficiency leads to severe feature ambiguity and sparse reconstruction voids in large-scale infrastructure scenes. Furthermore, existing deep learning descriptors typically neglect explicit spatial attributes and suffer from the computational burden of quadratic-complexity attention mechanisms, hindering deployment on edge devices. Building on the FeatureBooster-style descriptor enhancement paradigm, this study adapts a lightweight geometry-aware reconstruction framework to unmanned aerial vehicle infrastructure inspection. The methodology integrates a dual-stream descriptor enhancement model. Following the descriptor-boosting idea of combining local descriptors with geometric keypoint attributes, the adapted model embeds spatial attributes into the feature space to alleviate isomorphic ambiguity. Meanwhile, the modified cross-perception module replaces the original Attention-Free Transformer (AFT) -Simple setting with AFT-Full and incorporates SwiGLU to capture global context efficiently. Experiments on real-world datasets of complex interchanges, urban highways, and campus scenes validate that the adapted descriptor enhancement strategy effectively reduces point cloud voids and reduces reprojection error by approximately 17.3% compared with the original SIFT baseline on Dataset 1. Notably, this study empirically identifies an efficiency compensation phenomenon, wherein superior feature quality accelerates downstream geometric verification and optimization stages. Consequently, although feature enhancement introduces marginal overhead, the overall reconstruction time is reduced in certain datasets. This work provides an application-oriented adaptation and evaluation of FeatureBooster-style descriptor enhancement for geometry-consistent and computationally efficient infrastructure digitalization.

Keywords: Feature descriptor; Linear Attention; Structure-from-Motion; UAV photogrammetry; Weak texture.

INTRODUCTION

The unmanned aerial vehicle (UAV) serves as a flexible and efficient low-altitude remote sensing platform. Leveraging its high mobility and cost advantages, the UAV plays a critical role in emergency scenarios such as landslide investigation and monitoring (Sun *et al.*, 2024; Chen *et al.*, 2025) and flood monitoring or emergency response (Moussa *et al.*, 2024; Simantiris and Panagiotakis, 2024). Furthermore, it constitutes a core support for constructing digital twins of transport infrastructure (Yan *et al.*, 2023; Wu *et al.*, 2025). Benefiting from breakthroughs in sensor miniaturization, consumer UAVs equipped with high-resolution oblique photography modules and Real-Time Kinematic (RTK) positioning systems have achieved widespread adoption (Han and Han, 2024; Lee *et al.*, 2024). This proliferation significantly lowers the technical and economic barriers to acquiring massive high-fidelity geospatial data.

However, despite rapid advancements in frontend hardware, backend algorithms remain inadequate when confronting scenarios characterized by weak textures and isomorphic textures, such as highways. These conditions frequently lead to feature drift and reconstruction voids (Jiang *et al.*, 2020). Simultaneously, existing deep learning methods generally struggle to achieve stable matching in highly isomorphic regions due to a lack of explicit spatial geometric constraints (Ji *et al.*, 2023). Additionally, although the transformer architecture introduces necessary global perception capabilities, its

computational complexity, which grows quadratically with the number of feature points, fails to meet the requirements for efficient deployment on UAV edge devices (Sun *et al.*, 2021).

To address these challenges in UAV photogrammetric reconstruction, this paper adopts and adapts a FeatureBooster-style lightweight descriptor enhancement framework that combines explicit geometric embedding with global contextual information. First, this study constructs an explicit geometric embedding mechanism incorporating a feature descriptor self-mapping layer. This layer utilizes a multi-layer perceptron (MLP) to explicitly map spatial attributes of feature points onto a high-dimensional manifold and fuse them with visual features. This process endows descriptors with spatial location awareness. By adopting an Attention-Free Transformer (AFT) structure to reduce reliance on dense query-key attention computation, this module captures global context interactions among feature points with lower computational overhead. Finally, to resolve the weak-texture dilemma, we combine FastAP loss with descriptor enhancement loss to achieve end-to-end optimization, helping generate high-precision sparse point clouds with fewer geometric voids.

The main contributions of this paper are as follows. First, regarding geometric feature representation, this study adapts the FeatureBooster-style geometric descriptor enhancement mechanism to UAV SfM and evaluates its effectiveness in resolving matching ambiguity in isomorphic infrastructure textures. Second, concerning computational efficiency, we adopt attention-free context modeling inspired by FeatureBooster and modify the original AFT-Simple design by using AFT-Full and SwiGLU for UAV infrastructure reconstruction. Experimental results validate an efficiency compensation phenomenon, wherein the improvement in frontend feature purity reduces the computational overhead of backend geometric verification and optimization. Third, in the context of infrastructure digitalization, this study confirms that the proposed method effectively reduces geometric voids and reduces reprojection error. This constitutes a robust, low-cost, and high-precision solution for transport infrastructure.

LITERATURE REVIEW

Challenges in UAV-based Scene Reconstruction

With the rapid development of UAV aerial photography technology, the UAV serves as a primary data acquisition platform and has been widely applied in geospatial detection (Ke *et al.*, 2025). Structure-from-Motion (SfM) technology (Schönberger and Frahm, 2016) enables the recovery of camera poses and three-dimensional geometric structures using only multiple overlapping UAV images without auxiliary information (Westoby *et al.*, 2012; Jiang *et al.*, 2020), ultimately generating dense point clouds (Yao *et al.*, 2018). In recent years, high-frequency, high-resolution image acquisition based on UAVs has become a standard configuration for highway lifecycle maintenance (Koohmishi *et al.*, 2024). Although emerging rendering technologies such as Neural Radiance Fields (NeRF) and 3D Gaussian Splatting have achieved breakthrough progress in visual visualization (Kerbl *et al.*, 2023), point-based SfM and its derivative Multi-View Stereo (MVS) technologies continue to dominate engineering surveys and disease detection tasks requiring millimeter-level geometric accuracy due to their explicit geometric interpretability and engineering reliability (Jiang *et al.*, 2020).

Applying SfM technology to large-scale transport infrastructure, such as expressways and overpasses, presents specific challenges. Existing general reconstruction algorithms frequently encounter insufficient feature extraction or matching errors when processing such scenes due to uniform pavement materials or repetitive markings. Although Dusmanu *et al.* (2021) attempted to improve matching performance through cross-descriptor mapping, existing methods still face a precision-efficiency trade-off dilemma when confronting large-scale, high-dynamic UAV imagery.

Theoretical Evolution of Feature Matching

Image feature matching has transitioned from handcrafted designs to data-driven paradigms. Early algorithms such as SIFT (Lowe, 2004), ORB (Rublee *et al.*, 2011), and AKAZE (Alcantarilla *et al.*, 2011) dominated the field of 3D reconstruction by leveraging gradient statistical properties and real-time performance, while RootSIFT further optimized metric accuracy through kernel functions (Arandjelović and Zisserman, 2012). However, these methods are fundamentally limited by local patch extraction mechanisms. Due to the lack of spatial geometry constraints and global context perception, they often struggle in weak-texture and complex lighting scenarios.

With the development of deep learning, neural network-based feature extraction algorithms have demonstrated strong robustness. Yi *et al.* (2016) combined a spatial transformer network to achieve end-to-end training for feature point orientation estimation and descriptor extraction, significantly enhancing robustness under lighting and seasonal changes.

Subsequently, the SuperPoint algorithm proposed by DeTone *et al.* (2018) utilized a self-supervised framework to perform homographic adaptation between synthetic data and real images, substantially improving the repeatability of feature points.

Nevertheless, most such deep learning descriptors still primarily focus on visual appearance. Although some scholars have attempted to introduce geometric information for example through 3D point-cloud feature representation (He *et al.*, 2016) or evaluations of visual and geometric matching strategies (Ji *et al.*, 2023), existing visual descriptors remain unable to effectively distinguish feature points with extremely similar appearances but different positions when processing isomorphic structures with highly isomorphic textures, unless they incorporate explicit geometric embedding.

To further address the limitations of local isomorphism, resolving local ambiguity and introducing global perception has become an inevitable evolutionary direction in academia. Recent transformer-based matching methods have shown that global and cross-image contextual information is important for robust correspondence estimation. ASpanFormer introduces adaptive span attention to capture both global and local matching context (Chen *et al.*, 2022), MatchFormer interleaves self-attention and cross-attention for feature extraction and matching (Wang *et al.*, 2022), and LightGlue improves the efficiency of sparse local feature matching through an adaptive matching strategy (Lindenberger *et al.*, 2023). In addition, SOSNet improves descriptor learning by introducing second-order similarity constraints (Tian *et al.*, 2019). In a milestone work in this field, Sun *et al.* (2021) proposed the LoFTR algorithm, which introduced the architecture from Vaswani *et al.* (2017) into feature matching. By utilizing self-attention and cross-attention mechanisms to obtain a global receptive field, it effectively solved matching difficulties in low-texture areas. However, the computational burden of the standard transformer architecture, which exhibits quadratic complexity with resolution, makes it difficult to achieve real-time deployment on UAV edge devices with limited computing power. This approach of trading massive time costs for robustness severely restricts the reconstruction efficiency of large-scale scenes.

A closely related descriptor enhancement framework is FeatureBooster (Wang *et al.*, 2023), which improves existing local descriptors by integrating descriptor information with keypoint geometry through a lightweight neural network. However, existing studies have rarely examined the application of this descriptor-enhancement paradigm to UAV-based SfM reconstruction, especially for improving reconstruction completeness and spatial continuity in weak-texture infrastructure scenes.

Conceptual Framework

To address the aforementioned challenges, this study proposes a comprehensive conceptual framework (Fig. 1) aimed at enhancing algorithm adaptability and reconstruction stability in UAV scenarios with limited computing power. The framework comprises three hierarchical stages. First is the data acquisition layer. As the foundational layer, it is responsible for collecting multi-view high-resolution UAV images, providing raw visual data rich in geometric information for subsequent processing. Second is the descriptor-enhanced model.

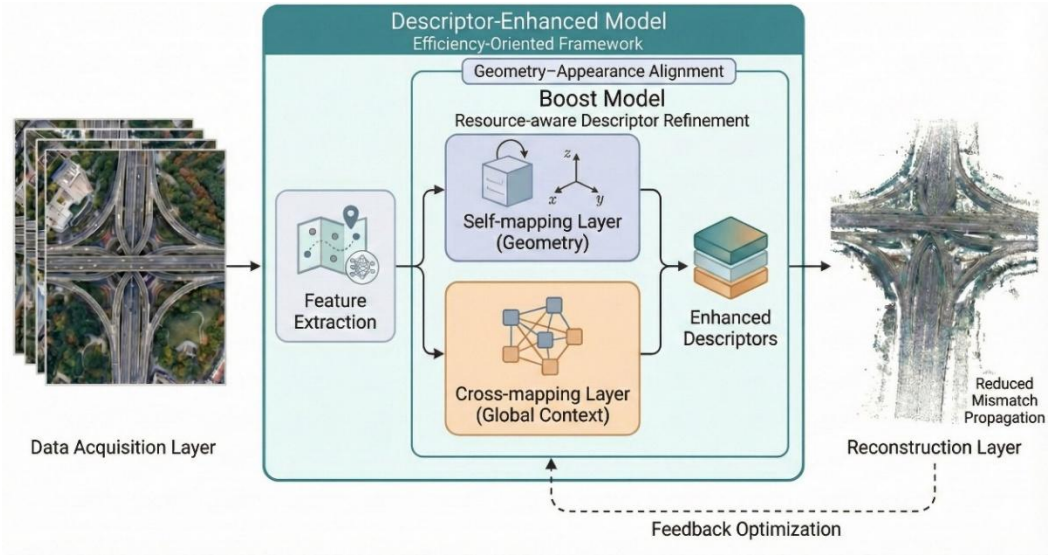

Fig. 1. *Conceptual framework of the proposed method.*

As the main adapted component, this layer follows the FeatureBooster-style descriptor boosting mechanism to address the geometry-appearance gap and the efficiency-perception gap in UAV SfM. Conceptually, the model integrates two parallel streams: a self-mapping layer (geometry) used to explicitly encode spatial attributes to supplement geometric constraints; and a cross-mapping layer (global context) used to capture global long-range dependencies to eliminate local ambiguity. Finally, there is the reconstruction layer. This layer utilizes the enhanced descriptors to drive the downstream 3D reconstruction pipeline. By inputting robust context-aware features, the framework facilitates the mitigation of feature drift and the reduction of potential geometric discontinuities, ultimately generating high-precision 3D models.

METHODOLOGY

Overall Framework and Incremental SfM

As illustrated in Fig. 2, this study adopts and adapts a FeatureBooster-style descriptor enhancement strategy within an incremental SfM sparse reconstruction framework. This framework aims to address matching ambiguity problems in weak-texture regions by embedding a lightweight descriptor enhancement model. The framework initiates with feature extraction and enhancement (Stage 1). The system first extracts raw SIFT (Lowe, 2004) or ORB (Rublee *et al.*, 2011) features from multi-view UAV images. Subsequently, these features are input into the proposed enhancement model (indicated by the red dashed box in Fig. 2). By explicitly embedding geometric constraints and global context, the model outputs enhanced descriptors, thereby significantly improving feature discriminability in complex scenes.

The process then proceeds to correspondence search and verification (Stage 2). This stage utilizes enhanced descriptors for correspondence search and employs geometric verification (RANSAC) to effectively suppress mismatches in weak-texture regions and reject outliers (Zhang *et al.*, 2025). Finally, in Stage 3 (Incremental Reconstruction), the system initiates the reconstruction loop based on scene graph initialization. It iteratively executes image registration and triangulation while jointly optimizing camera poses and 3D point coordinates via bundle adjustment. This process ultimately generates a geometrically consistent, high-precision sparse point cloud.

Geometric Self-Mapping Layer

Traditional feature descriptors rely solely on the statistical properties of local image gradients. Consequently, they often falter when processing isomorphic textures, such as identical lane lines or similar asphalt pavements (Morel and Yu, 2009). To introduce spatial constraints, this study follows the geometric fusion design used in FeatureBooster and adapts the first-stage self-mapping layer to the UAV SfM pipeline.

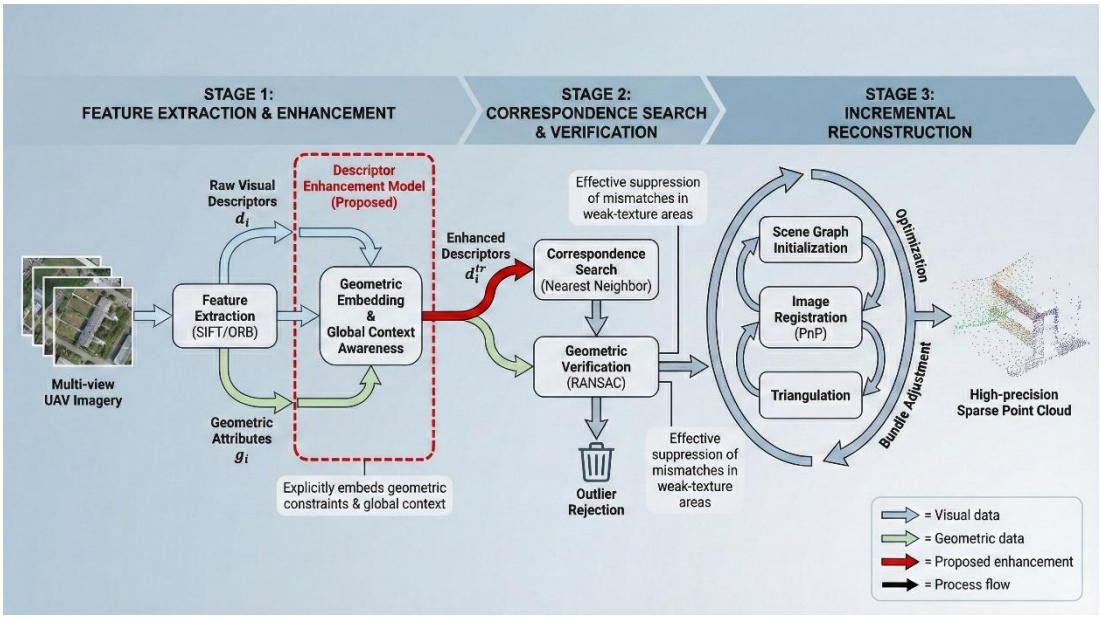

Fig. 2. *Incremental SfM Sparse Reconstruction Model Framework.*

As illustrated in Fig. 3, the model adopts a dual-stream parallel processing architecture. Let $d_i \in \mathbb{R}^D$ denote the original local descriptor of the i-th keypoint, and let $g_i = (x_i, y_i, \sigma_i, \theta_i) \in \mathbb{R}^4$ denote its geometric attribute vector, where x_i and y_i are normalized image coordinates, σ_i is the feature scale, and θ_i is the dominant orientation. The input stream on the left receives two types of heterogeneous data: the yellow channel inputs raw visual descriptors, while the blue channel inputs the corresponding geometric attribute vector.

Subsequently, the two inputs are processed via independent MLPs. The MLP serves as a universal function approximator, mapping the low-dimensional geometric Euclidean space to the same high-dimensional semantic manifold space as the visual descriptors (Hornik *et al.*, 1989).

Finally, in the feature fusion stage, the transformed geometric features and visual features undergo element-wise summation. As indicated by the red module in the Fig. 3, this operation achieves the explicit coupling of geometric priors and visual information, generating the intermediate descriptor d_i^s.

The mathematical expression of the self-mapping process is as follows:

$$d_i^s = \Phi_d(d_i) + \Phi_g(g_i) \quad (1)$$

where $\Phi_d(\cdot)$ and $\Phi_g(\cdot)$ denote the visual descriptor projection MLP and the geometric attribute projection MLP, respectively. Both projections map their inputs into the same D'-dimensional feature space, so that element-wise summation can be performed.

Lightweight Cross-Perception Network

Limitations of Standard Attention

While the self-mapping layer enhances the discriminability of individual feature points, it processes each point independently, neglecting co-visibility relationships and the global context. To capture these long-range dependencies, transformer architectures are widely applied (Han *et al.*, 2022). However, the standard multi-head attention (MHA) mechanism suffers from severe efficiency bottlenecks.

In a standard MHA layer, the input feature matrix $X \in \mathbb{R}^{N\times D}$ is projected into h attention heads through linear transformations, where N is the number of feature points and D is the descriptor dimension:

$$\begin{cases} Q_h = XW_h^Q \\ K_h = XW_h^K \\ V_h = XW_h^V \end{cases} \quad (2)$$

The attention output for the h-th head is calculated as follows:

$$Attention(Q_h, K_h, V_h) = Softmax\left(\frac{Q_h K_h^T}{\sqrt{D_k}}\right)V_h \quad (3)$$

Where D_k denotes the dimension of the key vector in each attention head. The core computational bottleneck lies in constructing the $N\times N$ attention matrix $Q_h K_h^T$, which leads to $O(N^2)$ time and memory complexity. Given that the number of feature points N in high-resolution UAV imagery typically reaches the order of $10^4 - 10^5$, this quadratic complexity leads to

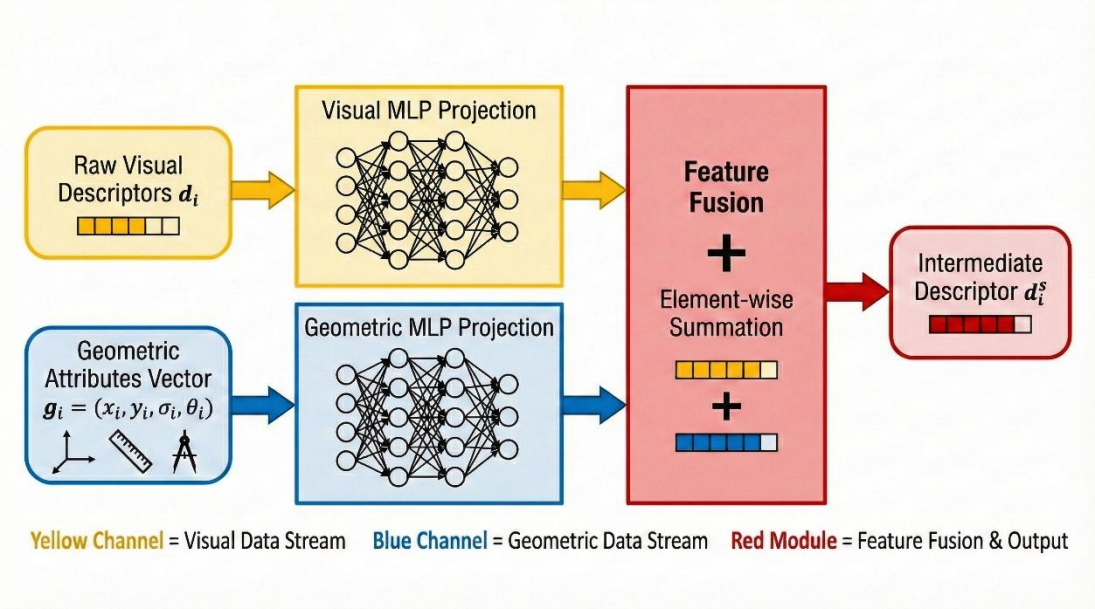

Fig. 3. *Descriptor Enhancement Model.*

memory explosion and fails to meet the requirements for rapid reconstruction on edge devices.

Attention-Free Transformer (AFT-Full)

To reduce computational costs while maintaining global perception, this study adopts attention-free context modeling inspired by FeatureBooster and replaces the original AFT-Simple setting with AFT-Full. As illustrated in Fig. 4, the AFT module reconstructs the information interaction method. First, the left path shows that the query vector Q is mapped to the (0,1) interval via the sigmoid activation function σ_q, serving as a gating mechanism. Second, the middle path introduces a position biases fusion mechanism, combining the key vector K with learned pairwise position biases w and weighting them with the value vector V in the exponential space.

Finally, following the attention-free modeling idea, AFT substitutes the computationally expensive matrix dot product found in traditional MHA with efficient element-wise multiplication ($\odot$), thereby reducing the dependence on explicit dense query-key attention computation. In this paper, the use of AFT-Full is treated as a modification of the FeatureBooster-style context module rather than as a newly introduced attention mechanism.

Compared with standard dot-product attention, AFT avoids explicitly computing the dense query-key attention matrix QK^T. With the low-rank decomposition of the pairwise position bias, the storage cost of the position-bias term is reduced from $O(N^2)$to $O(Nd')$, which is linear with respect to N when d' is fixed, where d' is the reduced embedding dimension.

$$f_i(X) = \sigma_q(Q_i) \odot \frac{\sum_{j-1}^{N} \exp(K_j + w_{i,j}) \odot V_j}{\sum_{j-1}^{N} \exp(K_j + w_{i,j})} \tag{4}$$

Where $w_{i,j}$ represents the learnable pairwise position bias between the i-th and j-th feature points. Q_i denotes the query representation of the i-th feature point, while K_j and V_j denote the key and value representations of the j-th feature point.

$$w_{i,j} = u_i^T v_j \tag{5}$$

Where u_i , v_j $\in \mathbb{R}^{d'}$ are low-rank learnable embeddings.

Feedforward Network Optimization (SwiGLU)

Following the AFT layer, to enhance the model's capability to fit complex non-linear relationships, this study improves the feedforward network (FFN). This study modifies the feedforward network by incorporating the SwiGLU activation function (Shazeer, 2020). This constitutes a variant structure combining Swish activation with a gated linear unit (GLU). As shown in Fig. 5, the SwiGLU module comprises a dual-path projection: the input vector is first split, wherein the gated generation path passes through transformation W and the Swish activation function to generate a non-monotonic gating signal; simultaneously, the information pass path preserves linear features via transformation V. Finally, the two signals undergo element-wise multiplication ($\otimes$) and are output via W_2.

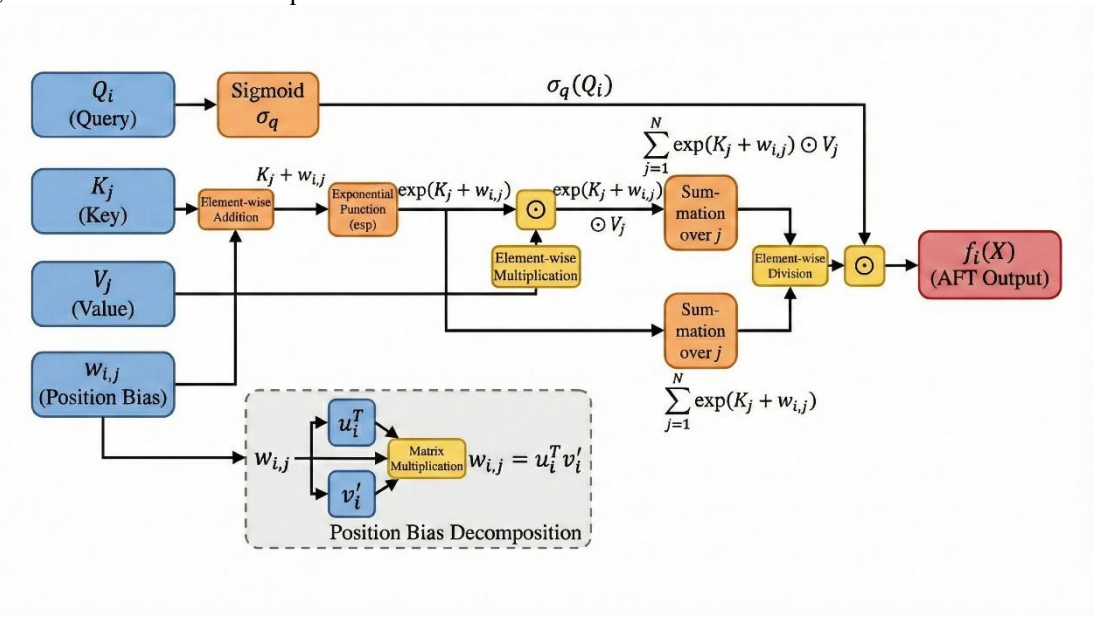

Fig. 4. *AFT Model.*

This mechanism allows the network to dynamically adjust the information throughput rate based on context, providing stronger feature selection capabilities compared to ReLU.

The standard FFN structure is:

$$FFN(x, W_1, W_2, b_1, b_2) = W_2[ReLU(W_1 x + b_1)] + b_2 \quad (6)$$

The improved FFN formula is as follows:

$$FFN_{SwiGLU} = [\mathrm{Swish}_\beta(\mathrm{xW}_1) \otimes \mathrm{xV}]\mathrm{W}_2 \quad (7)$$

Where $\mathrm{Swish}_\beta(z) = z \cdot \mathrm{sigmoid}(\beta z)$, W_1, V, and W_2 are trainable projection matrices, and β is the Swish activation parameter.

Loss Function Design

To drive the end-to-end training of the aforementioned network, this paper references the method of Cakir *et al.* (2019) to model the feature matching problem as a nearest-neighbor retrieval problem and adopts the Fast Average Precision (FastAP) loss function. Compared to triplet loss or contrastive loss, FastAP optimizes ranking metrics directly, making it more suitable for distinguishing extremely similar features.

Let M^+ and M^- denote the sets of matching and non-matching descriptor pairs, respectively. Let $z \in \Omega$ denote the Euclidean distance between a descriptor pair, where Ω is the distance domain. In the discretized FastAP formulation, Ω is quantized into a finite set Z of distance bins. Here, $F(z)$ denotes the cumulative distribution of all descriptor-pair distances up to bin z, $F(z|M^+)$ denotes the cumulative distribution of positive descriptor-pair distances up to bin z, and $P(M^+)$ denotes the prior proportion of positive pairs.

$$P(z) = \frac{F(z|M^+)P(M^+)}{F(z)} \quad (8)$$

$$R(z) = F(z|M^+) \quad (9)$$

From this, the Precision-Recall (PR) curve can be defined:

$$PR_z(\hat{\mathrm{d}}_\mathrm{i}) = \{[P(z), R(z)], z \in \mathrm{Z}\} \quad (10)$$

Average Precision (AP) is fundamentally the integral area under the PR curve (Cakir *et al.*, 2019):

$$P_{AP} = \int_{z\in\Omega} P(z)dR(z) \quad (11)$$

Since the integral is difficult to optimize directly, FastAP utilizes distance quantization techniques to discretize Z into a finite set of z values, thereby converting the integral into a differentiable summation form:

$$P_{FastAP} = \sum_{z\in Z} \frac{F(z|M^+)P(M^+)}{F(z)} P(z|M^+) \quad (12)$$

To maximize the retrieval accuracy of the enhanced descriptor $\hat{d}_i$, this study minimizes the FastAP loss:

$$L_{AP} = 1 - \frac{1}{N}\left[\sum_i^N P_{FastAP}(\hat{d}_i)\right] \quad (13)$$

Furthermore, following the boost-loss formulation used in FeatureBooster, this study uses a Boost loss constraint to encourage the AP value of the enhanced descriptor to be higher than that of the original descriptor d_i.

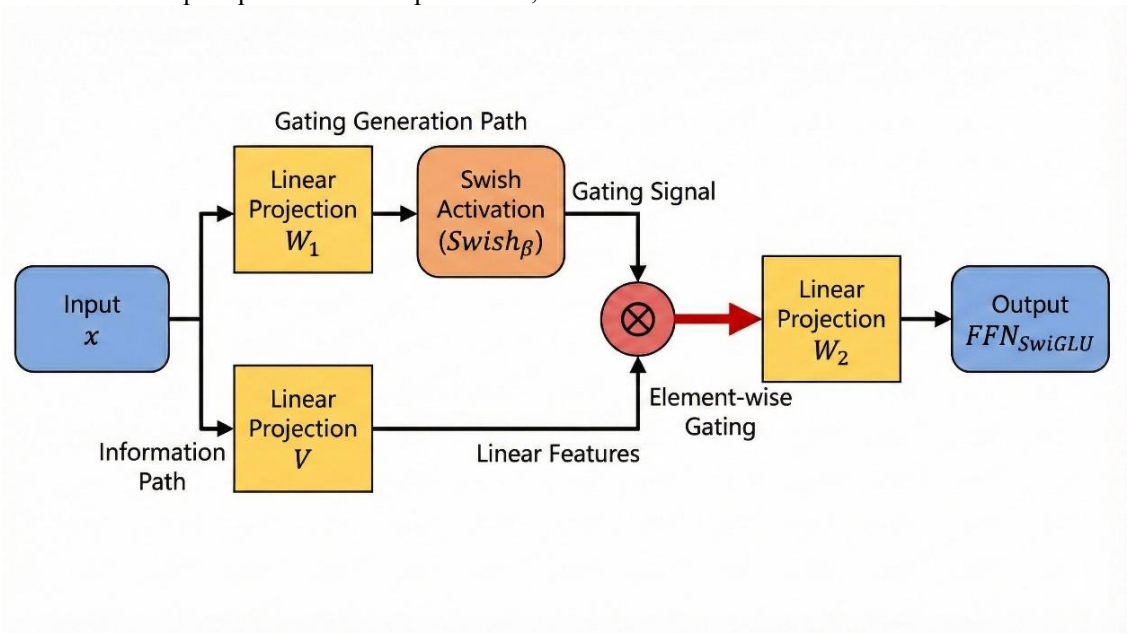

Fig. 5. *SwiGLU Model.*

$$L_{Boost} = \frac{1}{N}\sum_{i=1}^{N} max\left(0, \frac{P_{FastAP}(d_i)}{P_{FastAP}(\hat{d}_i)+\epsilon} - 1\right) \quad (14)$$

$\hat{d}_i$ denotes the final enhanced descriptor, and ϵ is a small constant to avoid division by zero. The final total loss function is the weighted sum of both components:

$$L = L_{AP} + \lambda L_{Boost} \quad (15)$$

Where λ is the balancing coefficient between the FastAP loss and the Boost loss. $P_{FastAP}(\hat{d}_i)$ and $P_{FastAP}(d_i)$ denote the FastAP scores of the enhanced descriptor and the original descriptor, respectively.

Methodological Workflow

As illustrated in Fig. 6, the algorithm proposed in this study constitutes an end-to-end closed-loop computational workflow. The workflow initiates with multi-modal data fusion (Step 0 and 1). The system extracts visual appearance data from raw images and geometric attributes data from keypoints in parallel. Through independent mapping layers, it completes the combine data streams operation within the latent space, generating fused features equipped with preliminary geometric perception. Subsequently, the data stream enters the global context understanding (Step 2) stage, utilizing the AFT module to capture long-range relationships and the SwiGLU module to refine features non-linearly, thereby achieving global enhancement and dynamic screening of features.

On this basis, the workflow diverges into two core paths. In the training path (red line), the system calculates the mixed loss based on the defined learning goal and executes backpropagation to update model parameters, establishing metric advantages in the feature space. In the inference path (green line), the trained model directly outputs the final enhanced descriptors, driving the improved 3D reconstruction pipeline (Step 3). Leveraging higher discriminability, the new features significantly optimize the matching and structure recovery process, ultimately generating a high-precision 3D point cloud. This realizes a complete closed loop from multi-modal input to 3D model construction.

EXPERIMENTS AND RESULTS

Experimental Setup and Datasets

To comprehensively evaluate the reconstruction performance of the proposed algorithm in scenarios characterized by geometric ambiguity and weak textures, this study integrated the proposed modules into an incremental SfM pipeline implemented with PyTorch-based components for feature learning. The experimental platform utilized a workstation architecture; specific hardware configurations and software environments are detailed in Table 1. All comparative experiments were conducted within a unified computing environment to ensure fair comparison across all methods.

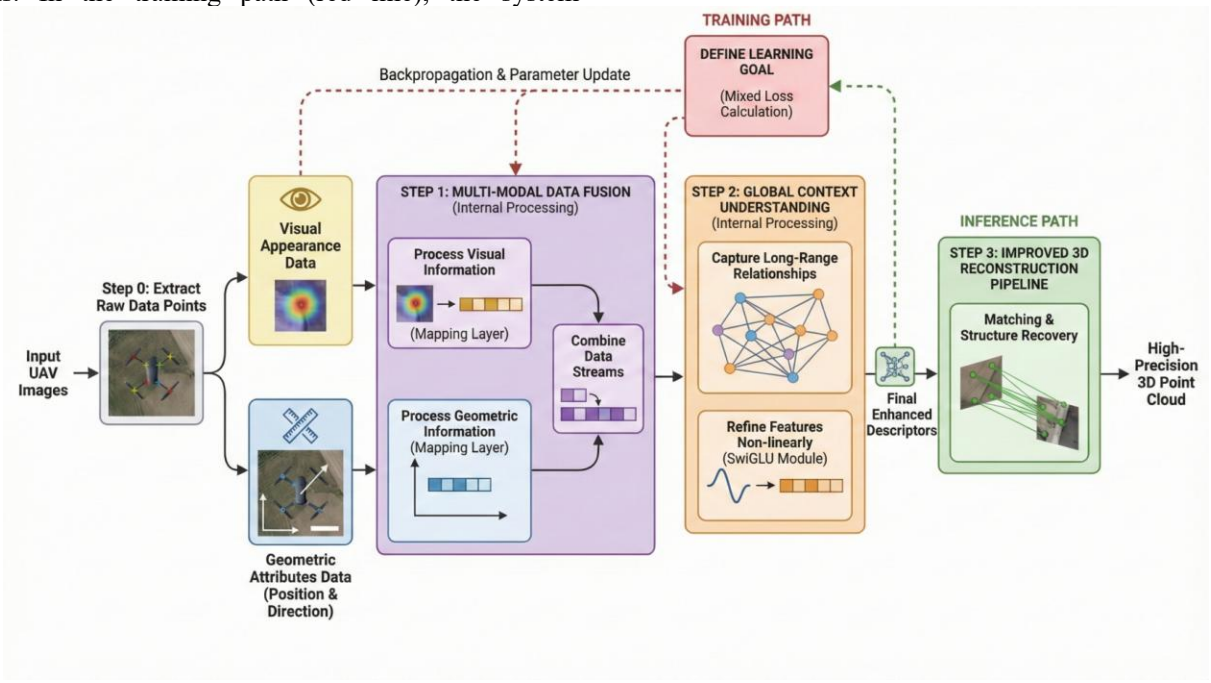

Fig. 6. *Methodological Workflow and Algorithm Overview.*

Table 1. *Hardware and Software Configuration*

Component	Specification
Processor (CPU)	Intel Core i7-13700KF @ 5.2 GHz (16 Cores)
Graphics Card (GPU)	NVIDIA GeForce RTX 4080 (16GB GDDR6X)
Memory (RAM)	32 GB DDR5
Framework	PyTorch 2.0.1 (CUDA 11.8)
OS Environment	Ubuntu 20.04 LTS / Windows 11

Following the method of Bu *et al.* (2016), this study employed two self-collected UAV oblique photography datasets and one public benchmark dataset for experimental evaluation. As illustrated in Fig. 7, these selected datasets represent varying degrees of scene complexity and texture characteristics, aiming to validate the robustness of the algorithm under different geometric configurations, texture distributions, and viewpoint variations. For the self-collected data, this study utilized a DJI-M3E UAV equipped with an RTK module, executing a five-direction flight route (one nadir + four oblique) to acquire high-resolution imagery (5280×3956 pixels). Significantly, all keyframes contain centimeter-level precision position and orientation system data, providing high-accuracy pose references for subsequent quantitative evaluation, rather than serving as direct supervision during reconstruction.

The specific characteristics of each dataset were defined to address distinct reconstruction challenges. Dataset 1 (urban highway, Fig. 7a) comprises 341 images with a shooting interval of 10 m. This scene constitutes a typical strip highway featuring significant repetitive patterns and weak-texture regions; it is specifically designed to test feature matching stability under conditions of visual feature scarcity. Dataset 2 (complex interchange, Fig. 7b) contains 396 images, covering multi-level overpasses and curved ramps. The drastic viewpoint changes and severe spatial occlusions in this dataset are intended to evaluate the spatial perception capability of the enhanced descriptors when processing complex geometric structures. Furthermore, Dataset 3 (NPU Campus, Fig. 7c) consists of 835 images representing a standard urban building complex. The introduction of this dataset aims to validate the generalization ability of the algorithm in a structurally regular yet texture-diverse urban environment.

Overall Sparse Reconstruction Performance

This section compares the adapted descriptor enhancement methods, SIFT+Boost (Ours) and ORB+Boost (Ours), with traditional handcrafted feature algorithms, SIFT and ORB (Lowe, 2004; Rublee *et al.*, 2011), earlier deep learning-based algorithms, SuperPoint and SOSNet (DeTone *et al.*, 2018; Tian *et al.*, 2019), and the state-of-the-art descriptor enhancement framework FeatureBooster (Wang *et al.*, 2023).

As shown in Table 2, the integration of the geometric perception enhancement module leads to consistent improvements across multiple quantitative metrics for the traditional algorithms. In Dataset 1, the reprojection error of SIFT+Boost decreases from 1.39 pixels (original SIFT) to 1.15 pixels, representing a reduction of 17.3%, while the number of sparse points increases by approximately 7.8%. In Dataset 2, the reprojection error of ORB+Boost decreases from 1.55 pixels to 1.38 pixels, accompanied by an increase in mean track length (from 2.51 to 2.79), indicating improved feature track stability across views. In the NPU_Central dataset, ORB+Boost records the highest number of sparse points (210,242) across the table, adding over 6,000 feature points compared to the original algorithm, while maintaining a reprojection error of 0.69 pixels, which is lower than that of the baseline ORB (0.88 pixels).

Fig. 7. *Datasets. (a) dataset 1; (b) dataset 2; (c) NPU_Central.*

Table 2. *Algorithm Performance Comparison*

Dataset	Algorithm	Sparse point ↑	Observation point ↑	Track length ↑	Reprojection error ↓
Dataset 1	SIFT	157,880	536,988	3.40	1.39
	ORB	34,383	80,355	2.33	1.49
	SuperPoint	169,722	594,048	3.50	1.32
	SOSNet	168,938	571,010	3.38	1.30
	FeatureBooster	169,500	585,210	3.45	1.25
	SIFT+Boost	170,117	610,878	3.59	1.15
	ORB+Boost	39,519	88,990	2.25	1.40
Dataset 2	SIFT	113,776	458,270	4.02	1.52
	ORB	47,028	118,316	2.51	1.55
	SuperPoint	134,512	583,577	4.34	1.39
	SOSNet	131,652	563,470	4.28	1.38
	FeatureBooster	135,240	590,115	4.30	1.36
	SIFT+Boost	139,817	602,335	4.31	1.35
	ORB+Boost	49,617	138,518	2.79	1.38
NPU_Central	SIFT	67,625	482,565	7.13	0.79
	ORB	203,783	700,618	3.43	0.88
	SuperPoint	39,555	263,680	6.66	1.25
	SOSNet	50,516	328,859	6.51	0.98
	FeatureBooster	55,210	350,110	6.70	0.85
	SIFT+Boost	76,439	574,838	7.52	0.68
	ORB+Boost	210,242	723,677	3.84	0.69

Compared to the deep learning-based SuperPoint, SOSNet, and the state-of-the-art descriptor enhancement framework FeatureBooster, the adapted methods demonstrate competitive and scenario-dependent performance across different datasets. FeatureBooster exhibits strong performance as a descriptor enhancement baseline, achieving reprojection errors of 1.25 pixels in Dataset 1 and 1.36 pixels in Dataset 2, outperforming SuperPoint and SOSNet in both cases. However, SIFT+Boost (Ours) records lower reprojection errors of 1.15 pixels and 1.35 pixels in Dataset 1 and Dataset 2, respectively. This result suggests that the adapted framework achieves competitive performance relative to FeatureBooster in the evaluated UAV infrastructure datasets. The observed improvement may be associated with the application-specific adaptation and the modified context modeling configuration (AFT-Full and SwiGLU), although a dedicated component-level ablation would be required to isolate the contribution of each modification. In the NPU_Central scenario, the point cloud density generated by each algorithm diverges significantly: SuperPoint generates 39,555 sparse points, whereas ORB+Boost and SIFT+Boost generate 210,242 and 76,439 points, respectively, indicating substantially higher reconstruction density in this scene.

Fig. 8 provides qualitative visualization results of the sparse reconstruction results before and after applying the FeatureBooster-style enhancement module. The red boxes indicate the same local regions across different methods before and after descriptor enhancement, so that the differences mainly reflect changes in point density, local completeness, and structural continuity rather than changes in viewpoint or scene content. As shown in Fig. 8, the original SIFT, ORB, SuperPoint, and SOSNet results still contain visible sparse regions and local discontinuities, especially in weak-texture road surfaces, interchange ramps, and regular campus structures. In contrast, the SIFT+Boost and ORB+Boost methods adapted and evaluated in this paper produce denser and more spatially continuous sparse reconstructions, more clearly recovering asphalt pavement details in Dataset 1 and complex interchange geometric structures in

Dataset 2, thereby reducing large geometric gaps. Notably, in the NPU_Central scenario (right column), certain deep learning methods (e.g., SuperPoint, SOSNet) exhibit local structural incompleteness in this dataset, presenting obvious structural deficiencies. Conversely, the adapted methods generate highly dense and structure-consistent sparse point clouds, suggesting improved matching robustness and reduced ambiguity in repetitive or weak-texture regions.

Ablation Study

To systematically evaluate the contribution of each module within the proposed framework, this study conducted an ablation study on Dataset 1. Table 3 presents the quantitative results following the gradual introduction of the self-mapping layer and the cross-mapping layer.

First, upon introducing the self-mapping layer, the model exhibited stable improvements across both feature extractors. For SIFT, the number of sparse points increased from a baseline of 157,880 to 162,305, while the reprojection error decreased from 1.39 px to 1.38 px. For ORB, although the reprojection error remained largely unchanged (1.49 px), the number of sparse points increased from 34,383 to 35,492. These results indicate that incorporating local geometric priors facilitates stability during the initial feature screening stage and retains more effective features without increasing error.

Second, the cross-mapping layer delivered more significant performance gains. When this module was enabled alone, the sparse points for SIFT increased to 167,549, and the reprojection error decreased to 1.35 px. For ORB, the sparse points rose to 38,751, and the error dropped from 1.49 px to 1.47 px. Compared to the self-mapping layer, the cross-mapping layer yielded more pronounced improvements for both feature extractors, suggesting that introducing cross-regional context information helps enhance the consistency and discriminability of feature descriptions.

When both modules were enabled to constitute the full model, the system achieved optimal performance across all metrics. The SIFT + Full Model obtained the highest number of sparse points (170,117) and the lowest reprojection error (1.15 px). The sparse points for the ORB + Full Model increased by approximately 14.9% compared to the baseline, and the reprojection error further decreased to 1.40 px. These results validate that local geometric constraints and global context modeling are complementary; their combination increases the effective number of feature matches while ensuring geometric consistency.

To intuitively demonstrate the changes in matching quality, Fig. 9 provides a visual comparison of feature matching results. It was observed that the baseline method exhibited obvious cross-matching in repetitive texture areas, whereas the full model generated more parallel and structurally consistent matching relationships. This qualitative result aligns with the quantitative metric trends, further confirming that the proposed modules effectively reduce mismatches and enhance matching stability.

Table 3. *Comparison of Ablation Experimental Data*

Method	Self-mapping Layer	Cross-mapping Layer	Sparse Points (↑)	Reprojection Error (px) (↓)
SIFT (Baseline)	-	-	157,880	1.39
	✓	-	162,305	1.38
	-	✓	167,549	1.35
SIFT + Full Model	✓	✓	170,117	1.15
ORB (Baseline)	-	-	34,383	1.49
	✓	-	35,492	1.49
	-	✓	38,751	1.47
ORB + Full Model	✓	✓	39,519	1.40

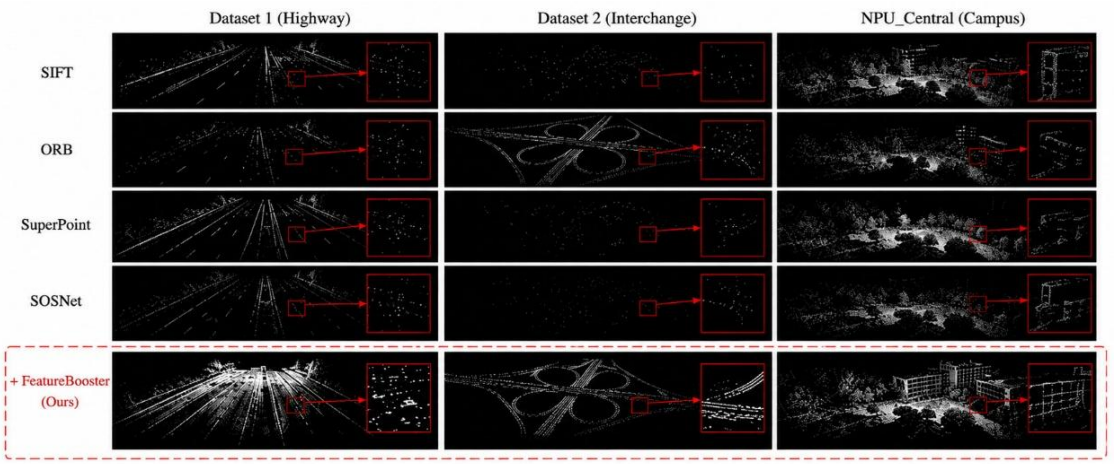

Fig. 8. *Qualitative comparison on three UAV SfM datasets.*

Computational Efficiency Analysis

Table 4 records the time consumption of the feature extraction stage for different algorithms across three datasets. Compared to the Baseline, in Dataset 1, the runtime for SIFT+Boost was 27.399 s, an increase of only about 1.6 s compared to the original SIFT (25.728 s); the time consumption for ORB+Boost was 28.589 s, with the increase controlled within 10%. After adding FeatureBooster as an additional descriptor enhancement baseline, its extraction time was 27.105 s in Dataset 1, 13.502 s in Dataset 2, and 12.015 s in NPU_Central, which is close to SIFT+Boost and confirms the lightweight nature of descriptor enhancement methods. This result indicates that the enhancement module is highly lightweight. Furthermore, compared to deep learning-based methods, the extraction efficiency of the proposed method in most scenarios was superior to or close to that of large models. For example, in Dataset 2, SIFT+Boost (13.809 s) was significantly faster than SuperPoint (19.788 s) and SOSNet (18.354 s), proving that the algorithm possesses good deployment potential on edge devices with limited computing resources.

Table 4. *Feature Extraction Running Time (Unit: s)*

Method	Dataset 1	Dataset 2	NPU_Central
SIFT	25.728	11.866	10.351
ORB	26.338	13.096	15.791
SuperPoint	27.578	19.788	14.559
SOSNet	26.354	18.354	13.854
FeatureBooster	27.105	13.502	12.015
SIFT+Boost	27.399	13.809	12.345
ORB+Boost	28.589	12.717	20.957

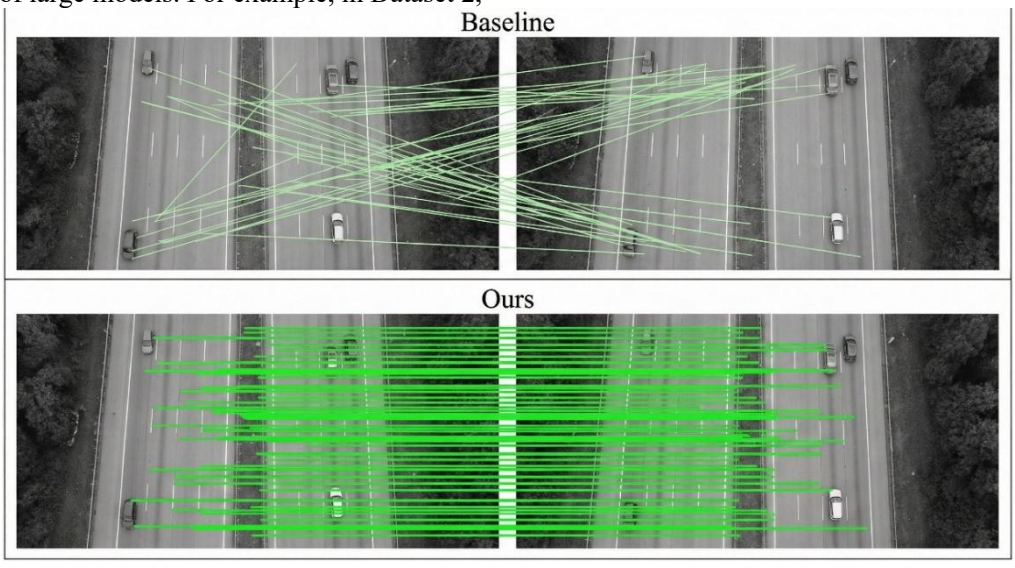

Fig. 9. *Visualization of feature matching improvements in the ablation study.*

Table 5 further reports the total reconstruction time, including feature matching, geometric verification, and bundle adjustment. In Dataset 1 and Dataset 2, the total reconstruction time for ORB+Boost (289.590 s and 414.135 s, respectively) was even lower than that of the original ORB algorithm (292.046 s and 521.194 s). After adding FeatureBooster as an additional baseline, its total reconstruction time was 405.120 s, 875.330 s, and 2850.410 s on Dataset 1, Dataset 2, and NPU_Central, respectively, which is close to the corresponding SIFT+Boost results. In the NPU_Central dataset, which has the largest data volume, although the total time consumption for SIFT+Boost was slightly higher than that of the original SIFT, its total time of 2836.873 s was far lower than that of the deep learning competitors SuperPoint (3885.129 s) and SOSNet (3186.541 s). This reflects its superior comprehensive computational efficiency when processing large-scale scenes.

Table 5. *Sparse Reconstruction Running Time (Unit: s)*

Method	Dataset 1	Dataset 2	NPU_Central
SIFT	382.538	839.299	2372.982
ORB	292.046	521.194	3637.606
SuperPoint	539.147	1047.862	3885.129
SOSNet	482.858	853.321	3186.541
FeatureBooster	405.120	875.330	2850.410
SIFT+Boost	401.469	867.480	2836.873
ORB+Boost	289.590	414.135	3563.741

DISCUSSION

This study aimed to address the dual challenges of geometric voids caused by weak textures and low computational efficiency at the edge during the construction of digital twins for large-scale transport infrastructure. The core findings indicate that adapting a FeatureBooster-style descriptor enhancement strategy to UAV SfM can help reduce matching ambiguity in isomorphic regions such as asphalt pavement and can also generate an efficiency compensation effect. Specifically, the backend geometric verification and optimization acceleration brought about by high-quality features sufficed to offset the computational cost introduced by frontend feature enhancement.

Mechanism Analysis and Performance Interpretation

Explicit Geometric Embedding for Disambiguating Isomorphic Textures

When processing highway pavements and repetitive markings, existing visual descriptors primarily face the challenge of severe degradation of discriminability. Although early deep learning explorations and SuperPoint (e.g., Yi *et al.*, 2016; DeTone *et al.*, 2018) significantly improved lighting robustness through end-to-end training, these methods fundamentally rely on statistical regularities of pixel intensity (appearance-based). As stated in the literature review, when confronting extreme isomorphic textures like asphalt pavement, the lack of explicit geometric constraints results in feature points being unable to be uniquely indexed in space. In contrast, the results of this study indicated that SIFT+Boost and ORB+Boost produced denser and more spatially continuous sparse reconstruction results in pavement areas in Dataset 1 and Dataset 2. This improvement is consistent with previous evaluations of visual and geometric matching strategies (Ji *et al.*, 2023), which indicate that geometric information can help alleviate matching ambiguity. It further suggests that, in highly repetitive infrastructure structures, explicit geometric indexing may improve descriptor stability and spatial discriminability.

These findings are also consistent with previous studies showing that local descriptor quality has a direct influence on feature matching, geometric verification, and subsequent SfM reconstruction performance (Lowe, 2004; DeTone *et al.*, 2018; Tian *et al.*, 2019). In particular, FeatureBooster (Wang *et al.*, 2023) demonstrated that existing local descriptors can be further improved by incorporating keypoint geometric information and lightweight neural refinement. The results of this study further support this view in UAV-based infrastructure reconstruction. Compared with the original descriptors, the FeatureBooster-style enhancement improves the density and continuity of sparse point clouds in weak-texture and repetitive scenes, indicating that descriptor enhancement is also effective in large-scale UAV SfM scenarios. Different from prior work that mainly evaluates descriptor enhancement in general visual matching tasks, this study focuses on UAV-based infrastructure SfM, where road surfaces, ramps, and regular building facades often contain weak texture and repeated structures.

This study further found that for highly repetitive engineering structures, relying solely on implicit context aggregation may be insufficient, and explicitly mapping spatial coordinates and dominant orientations to a high-dimensional manifold via the Self-mapping Layer can help improve descriptor separability. This encourages descriptors to become more distinguishable in the feature space for points with similar appearances but different positions. This finding complements He *et al.* (2016), which focused on 3D point cloud features while

not explicitly considering local visual descriptor coupling, and provides an application-oriented geometric enhancement strategy for alleviating SfM reconstruction problems in weak-texture regions mentioned by Jiang *et al.* (2020).

The Efficiency Compensation Effect: Non-Zero-Sum Game of Precision and Latency

Although milestone works represented by LoFTR (Sun *et al.*, 2021) resolved low-texture matching difficulties, the quadratic computational complexity of their introduced Transformer architecture (Vaswani *et al.*, 2017) with respect to image resolution makes it difficult to adapt to the high-frequency, full-lifecycle highway maintenance needs emphasized by Koohmishi *et al.* (2024). This study revealed a critical phenomenon: in the full-process reconstruction, the total time consumption of ORB+Boost was lower than that of the original ORB algorithm. This finding empirically indicates that in the SfM pipeline, precision and latency are not strictly a zero-sum relationship. Although the Boost module introduced a small amount of computation at the frontend, it significantly reduced the probability of mismatch propagation to subsequent stages by improving feature matching purity, thereby reducing invalid iterations in RANSAC geometric verification and the optimization burden of bundle adjustment. This indicates that in the multi-stage computational chain of SfM (Schönberger and Frahm, 2016), frontend feature quality exerts an amplified influence on backend optimization complexity. Its system-level effect likely occupies a higher weight in the overall runtime than being solely dominated by feature extraction speed. Therefore, this result can be interpreted as an efficiency compensation phenomenon, wherein the slight increase in frontend computation is partially or completely offset by the decrease in backend computational complexity at the system level. This finding provides empirical support for deploying high-precision reconstruction algorithms on UAV platforms with limited computing power, indicating that efficient engineering-grade reconstruction can be achieved by optimizing feature quality without relying on high-intensity GPU resources as required by NeRF (Kerbl *et al.*, 2023).

Linear Reconstruction of Global Context

Recent transformer-based matching methods, including ASpanFormer, MatchFormer, LoFTR, and LightGlue, have demonstrated the value of global and cross-image contextual information for robust correspondence estimation (Chen *et al.*, 2022; Wang *et al.*, 2022; Sun *et al.*, 2021; Lindenberger *et al.*, 2023). In addition, SOSNet improves descriptor learning by introducing second-order similarity constraints (Tian *et al.*, 2019). However, these methods also suggest that contextual robustness may be accompanied by increased computational cost, especially when applied to large-scale UAV SfM reconstruction.

Furthermore, the ablation experiments in this paper demonstrated that linear global perception realized through the AFT-Full module can improve reconstruction performance without relying on the full dense attention matrix. Unlike SuperPoint (DeTone *et al.*, 2018), which operates mainly on local visual feature representation, and distinct from the dense attention calculation of LoFTR (Sun *et al.*, 2021), the cross-mapping layer proposed in this paper utilized element-wise interaction to capture long-range contextual relationships of green belts and building facades. The improved structural continuity of the NPU_Central dataset in the experiment supports this observation: the combination of explicit geometry and global context enhanced the structural generalization capability of the model on unseen structures, rather than relying solely on appearance-driven memorization.

Research Implications

Theoretical Implications

First, this paper establishes the core status of explicit geometric embedding in weak-texture representation, correcting the purely vision-driven feature learning paradigm. Deep learning descriptors represented by SuperPoint (DeTone *et al.*, 2018) mainly rely on visual appearance representation and typically neglect explicit constraints on geometric attributes. Consequently, this appearance-dominated mechanism possesses inherent defects when facing isomorphic textures; that is, when local textures are highly consistent, simple implicit encoding loses spatial discriminability, leading to an inability to distinguish feature points with similar appearances but different physical positions. Although He *et al.* (2016) attempted to utilize 3D point cloud features and Ji *et al.* (2023) tried to aggregate geometric contexts, the former ignored local visual coupling, and the latter still struggled to effectively disambiguate in highly repetitive structures. The explicit geometric embedding mechanism proposed in this study indicates, from the levels of experimental results and mechanism analysis, that the explicit fusion of spatial geometric attributes such as position, scale, and direction with visual features constitutes a sufficient geometric condition for alleviating isomorphic texture ambiguity. This finding reveals that geometric attributes should serve as an

explicit index key rather than an implicit auxiliary feature.

Second, although existing global perception models such as LoFTR (Sun *et al.*, 2021) resolved matching difficulties in low-texture regions, the quadratic complexity of their adopted standard transformer (Vaswani *et al.*, 2017) architecture restricts their real-time application at the edge. The lightweight architecture based on AFT-Full and SwiGLU designed in this study suggests the potential usefulness of the linear attention mechanism in photogrammetry tasks from a theoretical level. This paper further demonstrates that for SfM tasks, the capture of global context may not always require the calculation of the full dense attention matrix; linear interaction via position bias can help construct a robust global topology. This finding suggests that in photogrammetry tasks, high-precision global perception does not necessarily depend on full quadratic complexity attention mechanisms. This finding provides a complementary perspective to previous transformer-based matching studies such as LoFTR (Sun *et al.*, 2021). While these studies have shown the importance of global contextual information for robust feature matching, their standard attention mechanisms usually introduce high computational costs. In contrast, the results of this study suggest that lightweight global context modeling based on AFT-Full can improve UAV SfM reconstruction while maintaining lower computational complexity. This provides a useful reference for designing real-time perception algorithms for onboard UAV processing in the future.

Practical Implications

First, this study provides a low-cost, high-precision technical path for the weak-texture dilemma, lowering the threshold for infrastructure digital twins. Although NeRF (Kerbl *et al.*, 2023) and 3D Gaussian Splatting have made progress in visualization, SfM remains dominant in engineering surveys requiring millimeter-level geometric accuracy. The experimental results of this study indicate that the algorithm effectively reduces point cloud voids and reduces reprojection error in asphalt pavement and complex interchange scenarios. This means that in practical engineering, utilizing existing consumer-grade UAVs combined with this algorithm can improve sparse reconstruction completeness without requiring major changes to the data acquisition platform. This breakthrough significantly reduces the marginal cost of transport infrastructure lifecycle maintenance and supports high-frequency digital archiving for transportation infrastructure digital twins (Yan *et al.*, 2023; Wu *et al.*, 2025), providing robust technical support for realizing high-frequency, automated pavement disease detection and digital archiving.

Second, this study validates the feasibility of algorithm compensating for hardware, enhancing the engineering value of existing UAV equipment. With the widespread application of UAVs in geospatial detection, traditional solutions often tend to upgrade frontend flight platforms or sensor hardware to cope with complex scenes (Ke *et al.*, 2025). This study proves that by optimizing the geometric perception capability of backend algorithms, it is possible to inversely compensate for the deficiencies of frontend sensors in texture capture. For generalized engineering scenarios not limited to highways, this hardware-software synergetic approach offers a highly cost-effective upgrade strategy: without replacing hardware, deploying enhanced algorithms alone can significantly improve the adaptability and data output quality of existing UAV operation systems, possessing broad industrial promotion value.

Limitations and Future Work

Our work has several limitations that suggest directions for future research. First, although the proposed framework improves UAV SfM reconstruction performance in the evaluated infrastructure scenes, its methodological novelty should be interpreted in an application-specific context. The general idea of descriptor enhancement is closely related to FeatureBooster (Wang *et al.*, 2023), which has shown that existing local descriptors can be strengthened by combining descriptor information with keypoint geometry. Therefore, the main contribution of this study lies not in proposing a completely new descriptor boosting paradigm, but in adapting and evaluating this paradigm for UAV-based infrastructure SfM, together with the AFT-Full and SwiGLU modification. In addition, the geometric embedding of this algorithm relies on the initial detection of frontend features. Under conditions of low signal-to-noise ratio such as night or heavy rain/fog, the degradation of underlying visual features may cause the geometric mapping to lose its physical basis. In the experiments, small fluctuations in performance metrics may occur due to variations in initial feature extraction, scene-specific texture distribution, and downstream geometric verification, although these variations remain within a controlled and acceptable range. Future work will improve evaluation robustness through repeated trials and cross-scene validation, while also exploring thermal infrared or LiDAR fusion to enhance reconstruction stability under low signal-to-noise conditions. Second, in unstructured

environments such as vegetation or fragmented terrain, the high geometric entropy of feature distribution limits the effectiveness of prior information. Subsequent research plans to develop an adaptive weighting mechanism to dynamically adjust the fusion ratio of geometric and visual streams based on scene semantics to enhance generalization capabilities in mixed scenes. Third, the current static topology assumption may be interfered with by transient geometric consistency when facing dense traffic flow, leading to residual dynamic artifacts. Future research will attempt to integrate lightweight semantic segmentation networks to pre-eliminate dynamic regions using semantic masks during the feature extraction stage, ensuring geometric purity of input data from the source.

CONCLUSION

As the construction of digital twins for transport infrastructure accelerates, high-resolution UAV photogrammetry has become a critical tool for lifecycle maintenance. However, the adaptability of existing algorithms remains limited by the dual bottlenecks of severe isomorphic texture challenges in large-scale highway scenes and low computational efficiency at the edge. This study constructs a comprehensive multi-view image dataset covering complex interchanges and urban roads and adapts a FeatureBooster-style dual-stream descriptor enhancement framework to UAV SfM reconstruction. The results confirm that utilizing geometric attributes as an explicit index key effectively reduces geometric voids in weak-texture pavements and repetitive markings. Furthermore, it achieves a breakthrough efficiency compensation effect, wherein high-quality feature input maximizes geometric precision while significantly shortening the full-process reconstruction time. This work provides a feasible low-cost solution for high-frequency automated infrastructure inspection and clarifies the practical value of adapting FeatureBooster-style descriptor enhancement to transport infrastructure digitalization. The general descriptor boosting architecture is attributed to prior work, while the contribution of this study lies in its UAV SfM adaptation, AFT-Full and SwiGLU modification, and infrastructure-scene evaluation.

Data Availability

The public dataset (NPU Campus) used in this study is available from the original source cited in the article. The self-collected datasets generated during the current study are available from the corresponding author on reasonable request.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Funding

This research was supported by the Research Team for Business Vocational Education and Industrial Economic Development at SIPIVT, China (Grant No. 20240103124).

REFERENCES

Alcantarilla PF, Solutions T (2011). Fast explicit diffusion for accelerated features in nonlinear scale spaces. IEEE Trans Pattern Anal Mach Intell 34(7):1281–98.

Arandjelović R, Zisserman A (2012). Three things everyone should know to improve object retrieval. https://doi.org/10.1109/CVPR.2012.6248018

Bu S, Zhao Y, Wan G, Liu Z (2016). Map2DFusion: Real-time incremental UAV image mosaicing based on monocular SLAM. Proc IEEE/RSJ Int Conf Intell Robots Syst, 4564–71. https://doi.org/10.1109/iros.2016.7759672

Cakir F, He K, Xia X, Kulis B, Sclaroff S (2019). Deep metric learning to rank. Proc IEEE/CVF Conf Comput Vis Pattern Recogn, 1861–70.

Chen H, Luo Z, Zhou L, Tian Y, Zhen M, Fang T, Quan L (2022). ASpanFormer: Detector-free image matching with adaptive span transformer. Eur Conf Comput Vis, 20–36. Cham: Springer Nature Switzerland.

Chen Y, Liu X, Zhu B, Zhu D, Zuo X, Li Q (2025). UAV image-based 3D reconstruction technology in landslide disasters: A review. Remote Sens 17(17):3117. https://doi.org/10.3390/rs17173117

DeTone D, Malisiewicz T, Rabinovich A (2018). SuperPoint: Self-supervised interest point detection and description. Proc IEEE Conf Comput Vis Pattern Recogn Workshops, 224–36.

Dusmanu M, Miksik O, Schönberger JL, Pollefeys M (2021). Cross-descriptor visual localization and mapping. Proc IEEE/CVF Int Conf Comput Vis, 6058–67.

Han S, Han D (2024). Enhancing direct georeferencing using real-time kinematic UAVs and structure from motion-based photogrammetry for large-scale infrastructure. Drones 8(12):736.

Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z, Tang Y, Xiao A, Xu C, Xu Y, Yang Z, Zhang Y, Tao D (2022). A survey on vision transformer. IEEE Trans Pattern Anal Mach Intell 45(1):1–1. https://doi.org/10.1109/tpami.2022.3152247

He L, Wang X, Zhang H (2016). M2DP: A novel 3D point cloud descriptor and its application in loop closure detection. Proc IEEE/RSJ Int Conf Intell Robots Syst. https://doi.org/10.1109/iros.2016.7759060

Hornik K, Stinchcombe M, White H (1989). Multilayer feedforward networks are universal approximators. Neural Netw 2(5):359–66. https://doi.org/10.1016/0893-6080(89)90020-8

Ji S, Zeng C, Zhang Y, Duan Y (2023). An evaluation of conventional and deep learning-based image-matching methods on diverse datasets. Photogramm Rec 38(182):137–59. https://doi.org/10.1111/phor.12445

Jiang S, Jiang C, Jiang W (2020). Efficient structure from motion for large-scale UAV images: A review and a comparison of SfM tools. ISPRS J Photogramm Remote Sens, 167, 230–51. https://doi.org/10.1016/j.isprsjprs.2020.04.016

Ke S, Yang K, Zhan C, Liao S, Ma Y, Shang H, Shen E, Li Z, Zhang Z, Chen Z (2025). Advances in earth observation using unmanned aerial vehicles: A bibliometric and content analysis, 2000–2024. Geocarto Int, 40(1). https://doi.org/10.1080/10106049.2025.2600779

Kerbl B, Kopanas G, Leimkühler T, Drettakis G (2023). 3D Gaussian splatting for real-time radiance field rendering. ACM Trans Graph 42(4):1–14. https://doi.org/10.1145/3592433

Koohmishi M, Kaewunruen S, Chang L, Guo Y (2024). Advancing railway track health monitoring: Integrating GPR, InSAR and machine learning for enhanced asset management. Autom Constr, 162, 105378. https://doi.org/10.1016/j.autcon.2024.105378

Lee E, Park S, Jang H, Choi W, Sohn HG (2024). Enhancement of low-cost UAV-based photogrammetric point cloud using MMS point cloud and oblique images for 3D urban reconstruction. Measurement, 226, 114158.

Lindenberger P, Sarlin PE, Pollefeys M (2023). LightGlue: Local feature matching at light speed. Proc IEEE/CVF Int Conf Comput Vis, 17627–38.

Lowe DG (2004). Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94

Morel JM, Yu G (2009). ASIFT: A new framework for fully affine invariant image comparison. SIAM J Imaging Sci 2(2):438–69. https://doi.org/10.1137/080732730

Moussa LG, Diaconu R, Watt MS, Muñoz E, Casado MR, Broadbent EN, Bruscolini M, Doaemo W, Mohan M (2024). UAVs as a tool for optimizing boat-supported flood evacuation operations. Drones 8(11):621. https://doi.org/10.3390/drones8110621

Rublee E, Rabaud V, Konolige K, Bradski G (2011). ORB: An efficient alternative to SIFT or SURF. https://doi.org/10.1109/ICCV.2011.6126544

Schönberger JL, Frahm JM (2016). Structure-from-Motion revisited. Proc IEEE Conf Comput Vis Pattern Recogn. https://doi.org/10.1109/CVPR.2016.445

Shazeer N (2020). GLU variants improve transformer. arXiv. https://doi.org/10.48550/arXiv.2002.05202

Simantiris G, Panagiotakis C (2024). Unsupervised color-based flood segmentation in UAV imagery. Remote Sens 16(12):2126. https://doi.org/10.3390/rs16122126

Sun J, Shen Z, Wang Y, Bao H, Zhou X (2021). LoFTR: Detector-free local feature matching with transformers. Proc IEEE/CVF Conf Comput Vis Pattern Recogn, 8922–31.

Sun J, Yuan G, Song L, Zhang H (2024). Unmanned aerial vehicles in landslide investigation and monitoring: A review. Drones 8(1):30. https://doi.org/10.3390/drones8010030

Tian Y, Yu X, Fan B, Wu F, Heijnen H, Balntas V (2019). SOSNet: Second order similarity regularization for local descriptor learning. Proc IEEE/CVF Conf Comput Vis Pattern Recogn, 11016–25.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones LJ, Gomez AN, Kaiser Ł, Polosukhin I (2017). Attention is all you need. Adv Neural Inf Process Syst, 30.

Wang Q, Zhang J, Yang K, Peng K, Stiefelhagen R (2022). MatchFormer: Interleaving attention in transformers for feature matching. Proc Asian Conf Comput Vis, 2746–62.

Wang X, Liu Z, Hu Y, Xi W, Yu W, Zou D (2023). FeatureBooster: Boosting feature descriptors with a lightweight neural network. Proc IEEE/CVF Conf Comput Vis Pattern Recogn, 7630–39.

Westoby MJ, Brasington J, Glasser NF, Hambrey MJ, Reynolds JM (2012). Structure-from-Motion photogrammetry: A low-cost, effective tool for geoscience applications. Geomorphology, 179, 300–14.

Wu D, Zheng A, Yu W, Cao H, Ling Q, Liu J, Zhou D (2025). Digital twin technology in transportation infrastructure: A comprehensive survey of current applications, challenges, and future directions. Appl Sci 15(4):1911. https://doi.org/10.3390/app15041911

Yan B, Yang F, Qiu S, Wang J, Cai B, Wang S, Hu W (2023). Digital twin in transportation infrastructure management: A systematic review. Intell Transp Infrastruct, 2, liad024.

Yao Y, Luo Z, Li S, Fang T, Quan L (2018). MVSNet: Depth inference for unstructured multi-view stereo. Proc Eur Conf Comput Vis, 767–83.

Yi KM, Trulls E, Lepetit V, Fua P (2016). LIFT: Learned invariant feature transform. Comput Vis ECCV, 467–83. https://doi.org/10.1007/978-3-319-46466-4_28

Zhang Y, Xue Y, Lan C, Xing X, Pang Y, Xu M (2025). Multisource oblique remote sensing image matching with affine-invariant features and geometric constraints. Int J Digit Earth, 19(1). https://doi.org/10.1080/17538947.2025.2600881