

STAR: Structure-aware representation learning for efficient and robust UAV-based geo-localization

Chengyue Wang^{1,2}, Bin Rao^{1,2}, Yanchen Guan^{1,2}, Jiaxun Zhang^{1,2}, Xingcheng Liu^{1,3}, Haicheng Liao^{1,3}, Zhenning Li^{1,2,3}✉

Cite this article: <https://doi.org/10.26599/COMMTR.2026.9640056>

ABSTRACT: Unmanned Aerial Vehicle (UAV)-based geo-localization supports autonomous navigation in GPS-denied or GPS-degraded environments by matching real-time UAV-view images with geo-referenced satellite imagery. A key challenge is to learn a discriminative retrieval space that remains robust to severe cross-view viewpoint discrepancies and temporal shifts. Although visual appearance provides fine-grained discriminative cues, it is sensitive to illumination, seasonal, and perspective variations. In contrast, geographic structures, such as road layouts and building footprints, offer more stable cues, but explicit structural modeling often increases computational cost. To balance robustness and efficiency, we propose **STAR**, a lightweight **ST**Structure-Aware Representation learning framework. STAR uses structural semantics as training-time geometric guidance for visual retrieval manifold learning. Specifically, a structural stream derives structure-aware representations from paired UAV and satellite images and constructs pairwise structural affinities to describe geographic layout relationships. These affinities regularize the discriminative retrieval space at the manifold level, enabling compact visual descriptors to retain structural awareness for robust cross-view matching. We further introduce UniT-1652, a benchmark for evaluating UAV-based geo-localization under temporal shifts. Extensive experiments on University-1652, SUES-200, and UniT-1652 show that STAR achieves state-of-the-art performance across multiple retrieval settings, with improved robustness to temporal and viewpoint variations and fewer parameters than existing methods.

KEYWORDS: Unmanned Aerial Vehicle; UAV-based Geo-Localization; Cross-view Geo-Localization; UAV Navigation

1 Introduction

Unmanned aerial vehicles (UAVs) are playing an increasingly important role in intelligent transportation systems, where transportation perception and operations are gradually extending from fixed ground-based sensing toward cooperative air-ground autonomy. Unlike ground vehicles and fixed roadside infrastructures, UAVs operate from controllable low-altitude viewpoints, which allows them to complement terrestrial sensing systems and acquire transportation-related information across broader, more dynamic, and less accessible regions (Z. Li et al., 2024; Yuan et al., 2025). Such capabilities have made UAVs valuable in a wide range of transportation applications, including traffic monitoring, emergency response, infrastructure inspection, and post-disaster assessment.

Localization is a fundamental capability for mobile robots, as it establishes the spatial reference required for perception, navigation, planning, and decision-making (Gao et al., 2025; Liao et al., 2024a). For UAVs deployed in various environments, this requirement is further complicated by GPS-denied or GPS-degraded conditions, which commonly arise in urban canyons, tunnels, underpasses, dense road networks surrounded by high-rise buildings, and post-disaster regions with damaged communication infrastructure (Chen et al., 2024). In such scenarios, accumulated drift from inertial measurements or visual

odometry can impair navigation reliability, making visual localization an essential complementary capability (Huang et al., 2024). UAV-based geo-localization addresses this problem by matching real-time UAV-view images with geo-referenced satellite-view images. The geographic metadata associated with the retrieved satellite images provide global localization evidence, which can support pose correction, drift recovery, route replanning, navigation-state verification, and higher-level mission decisions in UAV operation workflows (X. Wang et al., 2025). Technically, UAV-based geo-localization is commonly formulated as a cross-view image retrieval problem, where a UAV-view query image is matched against a geo-referenced satellite-view gallery. The objective is to learn a discriminative representation space in which geo-aligned UAV-satellite image pairs are embedded close to each other, while non-matching pairs are separated. Although recent deep learning methods have substantially improved cross-view representation learning for UAV-based geo-localization (Chen et al., 2025; T. Wang et al., 2021), their practical deployment remains constrained by two critical challenges.

The first challenge stems from *temporal shifts*. In real-world deployment, real-time UAV images are frequently matched against satellite imagery that may have been acquired months earlier (X. Wang et al., 2025; Zhang et al., 2024). This temporal lag introduces substantial appearance variations,

¹ State Key Laboratory of Internet of Things for Smart City, University of Macau, Macau 999078, China. ² Faculty of Engineering, University of Macau, Macau 999078, China.

³ Faculty of Information Science and Computing, University of Macau, Macau 999078, China.

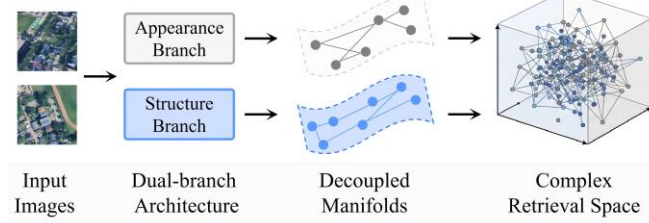
✉ Corresponding author. E-mail: zhenningli@um.edu.mo

Received: April 3, 2026; Revised: June 26, 2026; Accepted: September 24, 2026

© The Author(s) 2026. This is an open access article under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0, <http://creativecommons.org/licenses/by/4.0/>).

such as seasonal vegetation changes, snow coverage, and shifting illumination. Consequently, discriminative feature spaces established on datasets collected under synchronized timeframes struggle to generalize to the drastic distributional discrepancies caused by temporal shifts.

(a) Existing Approach



(b) Our Proposed STAR

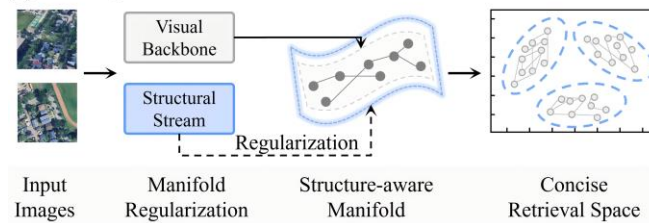


Fig. 1. Conceptual comparison of representation learning paradigms. (a) Existing structure-aware methods commonly rely on explicit appearance and structural modeling, resulting in decoupled manifolds and a complex retrieval space that imposes a heavy computational burden. (b) STAR introduces structural guidance to regularize the discriminative visual embedding space, shaping a compact structure-aware retrieval manifold for UAV-based geo-localization.

The second challenge arises from *viewpoint discrepancies*. UAV images are typically captured from oblique angles at varying altitudes, whereas satellite images provide a strict nadir perspective. These drastic perspective differences induce significant distortions in visual appearance, including texture stretching and object occlusion (Zhu et al., 2023a). Consequently, constructing a unified low-dimensional discriminative feature space is inherently difficult, as the severe domain discrepancy hinders the model from recognizing consistent and stable semantic features across both domains.

To address these challenges, UAV-based geo-localization has gradually evolved from appearance-dominated representation learning toward structure-aware representation learning. Early methods (Shi et al., 2020; T. Wang et al., 2021) mainly construct the retrieval space from visual appearance cues, such as color, texture, and local object patterns. Although these representations are effective under standard benchmark settings, they are inherently sensitive to temporal variations and cross-view perspective discrepancies, where the same geographic location may exhibit substantially different visual appearances. Structural semantics provide a more stable source of information for cross-view geo-localization. Geographic structures, such as road layouts, building footprints, and land-cover distributions, are generally less sensitive to seasonal changes, illumination variations, and moderate viewpoint differences

than local visual appearance. Motivated by this observation, recent methods (Q. Li et al., 2024; Zhang et al., 2023) have incorporated structural cues through multi-branch architectures to improve retrieval robustness. However, explicitly maintaining separate appearance and structural representations often increases model complexity, feature dimensionality, and inference cost, which is undesirable for UAV platforms with limited onboard computational resources. This raises an important question: *How can structural information be exploited to improve robustness while preserving a compact and efficient retrieval representation?*

To answer this question, we propose **STAR**, a lightweight **ST**tructure-**A**ware **R**epresentation learning framework for robust UAV-based geo-localization. Instead of treating structural semantics as an additional retrieval modality, STAR uses them as training-time geometric guidance to shape the discriminative retrieval manifold. Specifically, STAR introduces a structural stream to extract cross-view structural representations from paired UAV and satellite images. These representations are used to supervise structural consistency between matched cross-view samples and to construct structure-induced affinities that regularize the visual embedding space. As illustrated in Fig. 1, STAR transfers the stability of geographic structures into the learned visual descriptors, enabling the model to obtain a compact structure-aware representation for UAV↔Satellite retrieval. The main contributions are threefold:

- 1) We propose a structure-guided retrieval manifold learning framework for UAV-based geo-localization. STAR regularizes the discriminative visual embedding space with structural affinities, shaping a compact structure-aware manifold for robust UAV-based geo-localization.
- 2) We introduce UniT-1652, a benchmark built upon the University-1652 dataset that simulates deployment-time temporal shifts. UniT-1652 provides a controlled benchmark for evaluating robustness under deployment-relevant temporal appearance shifts.
- 3) We validate STAR on three benchmarks, University-1652, SUES-200, and UniT-1652, where it achieves state-of-the-art performance with fewer parameters. Ablation studies further confirm the critical role of structural awareness in enhancing robustness against temporal shifts and viewpoint discrepancies.

2 Related work

2.1. UAV-based geo-localization

UAV-based geo-localization aims to estimate the geographic location of a UAV by matching UAV-view images with geo-referenced satellite-view images. Existing deep learning methods mainly focus on constructing a shared discriminative embedding space for cross-view retrieval. Early approaches typically rely on appearance cues, such as color, texture, shape, and local object patterns, to align UAV and satellite images. CNN-based methods (Deuser et al., 2023; Guo et al., 2022) have been widely adopted for

extracting discriminative visual descriptors, while LPN (T. Wang et al., 2021) introduced a part-wise representation learning strategy to capture fine-grained contextual information. With the success of Transformer architectures in visual recognition, subsequent studies have explored attention-based models to improve global context modeling for UAV-based geo-localization (Zhang et al., 2023). For example, GeoFormer (Q. Li et al., 2024) introduced an efficient separable perceptron module to reduce the computational complexity of Transformer-based retrieval. Although these methods achieve competitive performance and efficiency on standard benchmarks, their representations are primarily driven by visual appearance. As a result, they can be sensitive to temporal appearance changes and viewpoint-induced geometric distortions, where the same geographic location may exhibit substantially different visual patterns across UAV and satellite views. To improve robustness, recent studies have introduced structural semantics into UAV-based geo-localization. Compared with local appearance, geographic structures such as road layouts, building footprints, and spatial scene layouts are relatively stable across time and viewpoints. Based on this observation, structure-aware methods incorporate semantic parsing, spatial layout modeling, attention-based structural reasoning, or multi-branch feature fusion to enhance cross-view representation learning. For instance, GeoDTR+ (Zhang et al., 2024) explicitly employed a geometric branch to extract layout information and fuse it with visual features. Other works have used Transformer-based modules to model spatial correlations across views (Zhang et al., 2023), or multi-head attention mechanisms to capture global structural relationships between UAV and satellite images (Wang et al., 2022). These methods demonstrate that structural semantics can provide complementary information for handling viewpoint and appearance variations.

Nevertheless, many existing structure-aware methods rely on explicit structural branches, multi-modal fusion, or high-dimensional joint representations to incorporate structural information. While effective for improving retrieval accuracy, such designs may increase model complexity and computational cost, which limits their suitability for resource-constrained UAV platforms. Unlike these approaches, STAR uses structural information to guide the learning of the retrieval representation itself. By deriving structural affinities from geographic semantics and using them to regularize the discriminative visual embedding space, STAR aims to learn a compact structure-aware retrieval manifold for efficient and robust geo-localization.

2.2. Learning using privileged information

Learning using privileged information (LUPI) refers to a learning paradigm in which additional information is available during training but absent during inference (Vapnik and Vashist, 2009; Yang et al., 2017). Rather than referring to a specific algorithm, LUPI provides

a general framework for transferring training-only knowledge to the deployed model. The goal is to improve generalization, robustness, or learning efficiency without increasing the computational cost of the deployed model (Alehdaghi et al., 2025). Existing LUPI-related methods can be organized according to the role that privileged information plays in the learning objective. The first category performs output-level knowledge transfer through teacher-student distillation. In this paradigm, a teacher model with richer inputs or stronger capacity provides privileged information, such as predictions, features, attention maps, or relational outputs, which are transferred to a compact student through feature-alignment loss functions. For example, Liao et al. (2024b) proposed HLTP++, a teacher-student framework for autonomous driving trajectory prediction. The second category introduces privileged information as auxiliary supervision (Liu et al., 2018; H. Wu et al., 2022). Under this formulation, training-only labels, semantic attributes, auxiliary modalities, or explanatory annotations are used to construct additional learning objectives that regularize the representation learned for the primary task. For example, Liu et al. (2019) proposed a hierarchical encoder with auxiliary supervision for table-to-text generation, where auxiliary sequence labeling, text auto-encoding, and multi-label classification tasks are jointly optimized to improve the semantic representation of structured tables. The third category incorporates privileged information at the constraint or optimization level. In this line of research, privileged features are used to characterize sample-dependent uncertainty, reliability, or difficulty, and are further embedded into the optimization objective through slack variables, margins, or instance-adaptive constraints. A representative example is SVM+ (Vapnik and Vashist, 2009), which introduces a correcting function in the privileged feature space to estimate the slack variables of standard SVMs, thereby allowing privileged information to guide the formation of the decision boundary. Following this formulation, subsequent studies have extended LUPI to a wide range of learning problems, including ranking, clustering, Gaussian processes, multi-task and multi-class classification, domain adaptation, and metric learning (Z. Wu et al., 2022; You et al., 2017). These methods prove that privileged information can be exploited to modulate the optimization geometry of the primary task, providing a principled mechanism for improving learning behavior when additional training-time information is available.

LUPI provides a promising perspective for improving the robustness of UAV-based geo-localization without increasing inference complexity. From this perspective, structural information can serve as privileged guidance to shape the visual retrieval space during training. Specifically, STAR converts structure-aware representations into pairwise structural affinities and imposes them on the discriminative retrieval space as manifold regularization, rather than using them as an auxiliary inference.

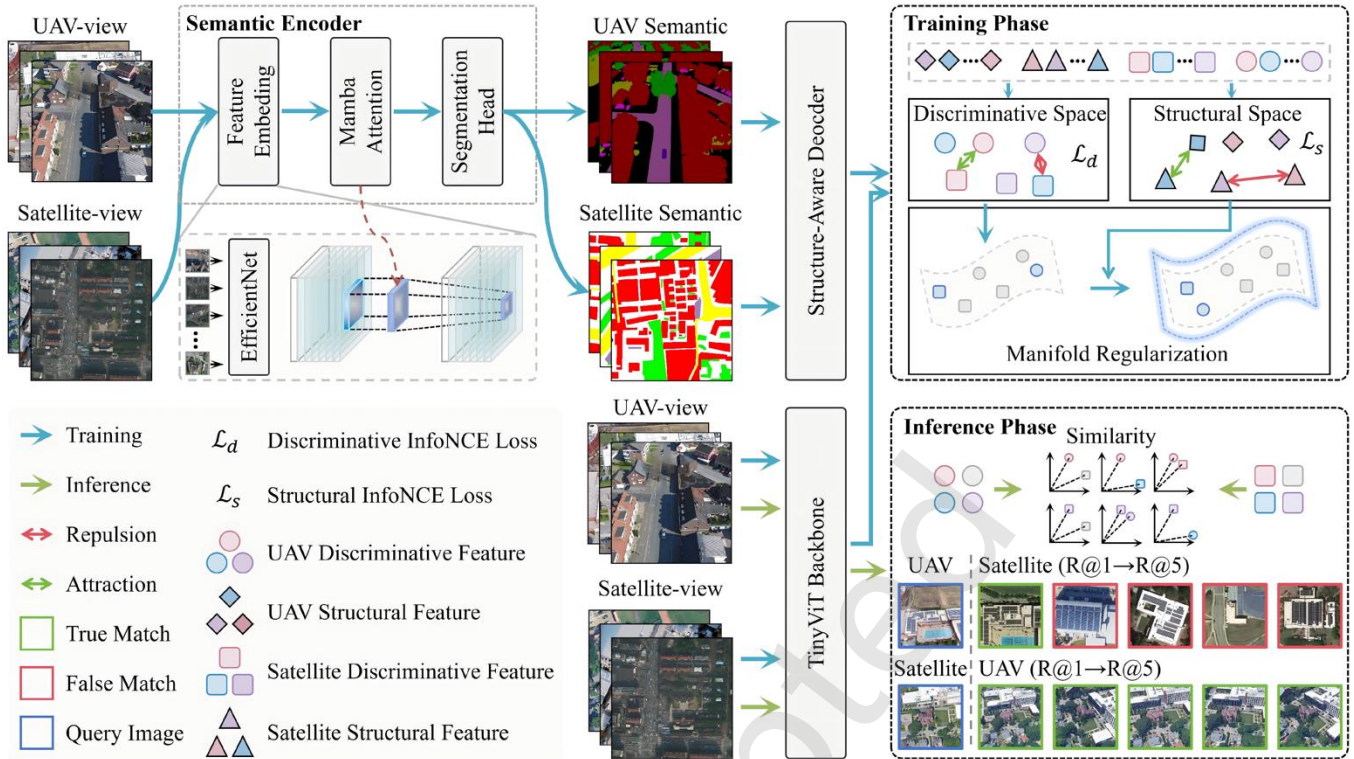


Fig. 2. The overall framework of the proposed lightweight structure-aware representation learning model, termed STAR. During training, STAR leverages a structural stream to guide the TinyViT backbone, enabling it to learn structure-aware discriminative features that are robust to temporal shifts and viewpoint discrepancies.

3 Methodology

3.1. Problem formulation

UAV-based geo-localization plays a vital role in aerial navigation and surveillance. This task is formulated as a bidirectional cross-view image retrieval problem between UAV and satellite views (Zheng et al., 2020). The first sub-task is target localization, where the objective is to retrieve the GPS-tagged satellite-view image that is most similar to a given UAV-view image, thereby inferring the position of the target object. The second sub-task is UAV navigation, which uses a satellite-view image as a reference to retrieve the corresponding UAV-view image, enabling the UAV to navigate toward a designated target area. In practice, UAV and satellite images are encoded as discriminative features (Xia et al., 2024; Zhang et al., 2024), with cross-view similarity computed using cosine similarity.

3.2. Model overview

As illustrated in Fig. 2, the proposed STAR adopts an asymmetric learning strategy that treats structural information as privileged knowledge. Specifically, the framework integrates two parallel pathways during the training phase: (1) A visual backbone (implemented via TinyViT (K. Wu et al., 2022)) that extracts discriminative features F_d from raw images; (2) A structural stream, which comprises a Semantic Encoder to generate appearance-agnostic semantic maps M_s , and a Structure-Aware Decoder to transform these maps into invariant structural features F_s . To ensure efficiency, the Structural Stream is detached during inference, and STAR operates solely using the lightweight TinyViT backbone.

3.3. Semantic encoder

Rather than extracting structural features directly from images, which are often entangled with appearance cues, STAR employs a semantic encoder to generate semantic maps M_s as an intermediate representation, as outlined in Algorithm 1. This design is based on the intuition that structural understanding can be decomposed into two stages: recognizing the semantic entities present in the scene and then reasoning about their spatial organization. The semantic encoder therefore converts the input image into category-level scene semantics, which provide a stable basis for subsequent structure-aware representation learning.

3.3.1 Feature embedding

Given an input image $I \in \mathbb{R}^{3 \times H \times W}$, the semantic encoder first adopts an EfficientNet backbone (Tan and Le, 2019) to extract a hierarchy of multi-scale features $\{F_1, F_2, F_3, F_4\}$. Here, $F_i \in \mathbb{R}^{C_i \times H_i \times W_i}$ denotes the feature map at the i -th level. These features gradually increase in semantic abstraction while decreasing in spatial resolution, allowing the encoder to capture both local details and high-level semantic context.

To strengthen the interaction between low-level spatial details and high-level semantic representations, we construct enhanced encoder features through downsampled skip connections and feature recalibration (Hwang et al., 2024). Specifically, shallow- and middle-level features are first projected to deeper feature levels and fused by element-wise addition after resolution and channel alignment:

$$G_2 = F_2 + \Phi_D^{1 \rightarrow 2}(F_1), \quad G_3 = F_3 + \Phi_D^{2 \rightarrow 3}(F_2), \quad (1)$$

where $\Phi_D^{i \rightarrow j}(\cdot)$ denotes a downsampling projection from level i to level j . The fused features are then recalibrated by

level-specific feature recalibration modules:

$$\tilde{F}_2 = \Phi_R^{(2)}(G_2), \quad \tilde{F}_3 = \Phi_R^{(3)}(G_3). \quad (2)$$

Here, $\Phi_R^{(2)}$ and $\Phi_R^{(3)}$ denote recalibration modules at different feature levels.

Based on these recalibrated features, we obtain the enhanced representations $\hat{F}_3$ and $\hat{F}_4$:

$$\hat{F}_3 = \Phi_D^{2 \rightarrow 3}(\tilde{F}_2) + \tilde{F}_3, \quad \hat{F}_4 = F_4 + \Phi_D^{3 \rightarrow 4}(\tilde{F}_3). \quad (3)$$

Here, $\hat{F}_3$ integrates middle-level structural details, while $\hat{F}_4$ incorporates high-level semantic context. These enhanced features provide the basis for subsequent semantic decoding.

Algorithm 1. Semantic Encoder

Definitions: $\Phi_D^{i \rightarrow j}$: downsample projection from level i to level j for resolution and channel alignment; $\Phi_R^{(i)}$: feature recalibration module at level i ; Φ_C : convolutional projection; Φ_B : Mamba-based residual block; Φ_W : weighted fusion block; Φ_F : feature refinement head; Φ_I : bilinear interpolation; Φ_S : segmentation head. The symbol $+$ denotes element-wise addition after resolution and channel alignment.

Input: Image $I \in \mathbb{R}^{3 \times H \times W}$

Output: Semantic map $M_s \in \mathbb{R}^{K \times H \times W}$

1. Extract multi-scale encoder features:
2. $\{F_1, F_2, F_3, F_4\} \leftarrow \text{EfficientNet}(I)$
3. Construct enhanced encoder features:
4. $G_2 \leftarrow F_2 + \Phi_D^{1 \rightarrow 2}(F_1)$
5. $G_3 \leftarrow F_3 + \Phi_D^{2 \rightarrow 3}(F_2)$
6. $\tilde{F}_2 \leftarrow \Phi_R^{(2)}(G_2)$
7. $\tilde{F}_3 \leftarrow \Phi_R^{(3)}(G_3)$
8. $\hat{F}_3 \leftarrow \Phi_D^{2 \rightarrow 3}(\tilde{F}_2) + \tilde{F}_3$
9. $\hat{F}_4 \leftarrow F_4 + \Phi_D^{3 \rightarrow 4}(\tilde{F}_3)$
10. Initialize decoder feature:
11. $F_4^d \leftarrow \Phi_B(\Phi_C(\hat{F}_4))$
12. Progressive semantic decoding:
13. $F_3^d \leftarrow \Phi_B(\Phi_W(F_4^d, \hat{F}_3))$
14. $F_2^d \leftarrow \Phi_B(\Phi_W(F_3^d, F_2))$
15. $F_m \leftarrow \Phi_F(F_2^d, F_1)$
16. Generate semantic map:
17. $M_s \leftarrow \Phi_I(\Phi_S(F_m))$
18. Return M_s

3.3.2 Mamba attention

To enhance the structural discriminability of semantic representations, we introduce Mamba-based residual blocks to model long-range contextual relationships among scene components. Unlike standard convolutional operators that mainly focus on local neighborhoods, the Mamba module (Gu and Dao, 2023) treats spatial positions as a sequence and efficiently captures global dependencies across the feature map.

Starting from the enhanced high-level feature $\hat{F}_4$, the decoder first projects it into a unified feature space and applies a Mamba-based residual block. For a decoder feature F , the block updates it as:

$$F' = F + \Phi_M(\text{Norm}(F)), \quad (4)$$

$$\Phi_B(F) = F' + \Phi_P(\text{Norm}(F')), \quad (5)$$

where $\Phi_M(\cdot)$ denotes the Mamba module for global contextual modeling, and $\Phi_P(\cdot)$ denotes a point-wise feed-

forward transformation.

The decoder then progressively integrates multi-scale encoder features through weighted fusion. Given the current decoder feature F and a skip feature S , the weighted fusion block is formulated as:

$$\Phi_W(F, S) = \Phi_{\text{post}}(\omega_s \cdot \Phi_{\text{skip}}(S) + \omega_d \cdot \Phi_I(F)), \quad (6)$$

where $\Phi_I(\cdot)$ upsamples the decoder feature, $\Phi_{\text{skip}}(\cdot)$ aligns the channel dimension of the skip feature, and $\Phi_{\text{post}}(\cdot)$ refines the fused representation. The coefficients ω_s and ω_d are normalized learnable weights that adaptively balance the contributions of the skip feature and the decoder feature. Following the top-down decoding order, the decoder first fuses the high-level decoder feature with $\hat{F}_3$, and then fuses the resulting feature with F_2 . A Mamba-based residual block is applied after each fusion stage.

3.3.3 Segmentation head

At the final decoding stage, a feature refinement head integrates the decoded representation with the highest-resolution encoder feature F_1 :

$$F_m = \Phi_F(F, F_1). \quad (7)$$

In the implementation, $\Phi_F(\cdot)$ first upsamples the decoder feature and applies channel shuffle, then fuses it with the high-resolution encoder feature using normalized learnable weights. The fused feature is further refined through spatial and channel attention branches, followed by residual separable convolution. This stage improves local spatial consistency before semantic prediction. The refined feature F_m is then passed to a lightweight segmentation head to obtain the semantic logit map:

$$M_s = \Phi_I(\Phi_S(F_m)), \quad (8)$$

where $M_s \in \mathbb{R}^{K \times H \times W}$ and K denotes the number of semantic classes. Here, $\Phi_S(\cdot)$ consists of convolutional refinement, dropout regularization, and a class-wise projection layer, while $\Phi_I(\cdot)$ resizes the prediction to the original image resolution. During pretraining, an auxiliary segmentation head is attached to intermediate decoder features to provide additional multi-scale supervision.

3.4. Structure-aware decoder

To obtain an appearance-agnostic representation of geographic layout, we introduce a graph-based Structure-Aware Decoder that explicitly models spatial relations among semantic regions. Given the semantic map M_s produced by the semantic encoder, we first convert it into a semantic probability map by applying a softmax operation. The semantic probability map is then adaptively pooled into a $G \times G$ regular grid, where $G = 12$ in our implementation. Each grid cell is regarded as a graph node, resulting in $N = G^2$ nodes. For each node, we construct its feature by concatenating the local semantic distribution with its normalized two-dimensional spatial coordinate. The semantic distribution is obtained from the corresponding pooled grid cell, and the coordinate is normalized to the range $[-1, 1]$. Therefore, each node encodes both the semantic composition of a local region and its position in the global scene layout. The node feature is first projected into a hidden dimension of 64 by a linear layer.

We then construct a spatial graph over the grid nodes using 8-neighborhood connectivity. Each node is connected to its horizontal, vertical, and diagonal neighboring cells. The adjacency matrix is row-normalized and used to aggregate information from neighboring regions. Let $\mathbf{h}_i^{(\ell)}$ denote the feature of node i at the ℓ -th graph layer. The graph convolution operation is formulated as:

$$\bar{\mathbf{h}}_i^{(\ell)} = \sum_{j \in \mathcal{N}(i)} A_{ij} \mathbf{h}_j^{(\ell)}, \quad (9)$$

$$\mathbf{h}_i^{(\ell+1)} = \text{Dropout} \left(\phi \left(\text{LN}(\mathbf{W}_s^{(\ell)} \mathbf{h}_i^{(\ell)} + \mathbf{W}_n^{(\ell)} \bar{\mathbf{h}}_i^{(\ell)}) \right) + \mathbf{R}^{(\ell)} \mathbf{h}_i^{(\ell)} \right), \quad (10)$$

where $\mathcal{N}(i)$ denotes the neighboring nodes of node i , A_{ij} is the normalized adjacency weight, $\mathbf{W}_s^{(\ell)}$ and $\mathbf{W}_n^{(\ell)}$ are learnable transformations for the self-node feature and the aggregated neighboring feature, respectively, $\text{LN}(\cdot)$ denotes layer normalization, $\phi(\cdot)$ is the GELU activation function, and $\mathbf{R}^{(\ell)}$ is a residual projection.

Through graph message passing, semantic information is propagated across neighboring regions, allowing the decoder to model relational layout cues beyond isolated semantic categories. This design enables the structural representation to encode not only the presence of semantic entities, such as roads and buildings, but also their spatial continuity, adjacency, and relative arrangement.

After graph reasoning, the updated node features are reshaped back into a $G \times G$ spatial layout and summarized by adaptive average pooling to a 3×3 feature map. The pooled feature is flattened and projected into a 576-dimensional structural embedding $\mathbf{F}_s \in \mathbb{R}^{576}$, which has the same dimensionality as the discriminative retrieval feature $\mathbf{F}_d$. This structural embedding is used during training to provide structure-aware supervision and manifold regularization for learning robust UAV-based geo-localization representations.

3.5. TinyViT backbone

To extract informative discriminative features $\mathbf{F}_d$ from images, we adopt TinyViT (K. Wu et al., 2022). TinyViT integrates convolutional token embeddings and hierarchical transformer blocks, enabling it to process high-resolution inputs while maintaining efficiency. Specifically, TinyViT employs window-based attention mechanisms to localize computation, avoiding the quadratic complexity typically associated with full self-attention. In early stages, convolutional layers are used to capture fine-grained local textures, while subsequent transformer blocks progressively encode higher-level visual semantics. This design ensures that TinyViT remains both lightweight and expressive.

3.6. Optimization objectives

To facilitate robust geo-localization, STAR is optimized with a compound objective that jointly enhances instance-level discriminative alignment, structural semantic learning, and structure-guided manifold regularization. Formally, the overall objective is defined as:

$$\mathcal{L} = w_d \mathcal{L}_d + w_s \mathcal{L}_s + w_r \mathcal{L}_r, \quad (11)$$

where $\mathcal{L}_d$ optimizes the retrieval capability of the

discriminative visual stream, $\mathcal{L}_s$ supervises the auxiliary structural stream, and $\mathcal{L}_r$ imposes structure-guided manifold smoothness on the learned discriminative representation. The coefficients w_d , w_s , and w_r balance the contributions of the three terms.

3.6.1 Dual contrastive learning ($\mathcal{L}_d$ and $\mathcal{L}_s$)

We employ a symmetric InfoNCE loss (Deuser et al., 2023) for both the discriminative and structural streams. Let $\{(I_i^u, I_i^s)\}_{i=1}^B$ denote a mini-batch of B matched UAV-satellite image pairs. For the discriminative stream, given the embeddings $\mathbf{F}_{d,i}^u, \mathbf{F}_{d,i}^s \in \mathbb{R}^d$, the bidirectional contrastive loss is formulated as:

$$\mathcal{L}_d = \frac{1}{2} (\mathcal{L}_d^{u \rightarrow s} + \mathcal{L}_d^{s \rightarrow u}), \quad (12)$$

where the UAV-to-satellite alignment term is defined as:

$$\mathcal{L}_d^{u \rightarrow s} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp((\mathbf{F}_{d,i}^u)^\top \mathbf{F}_{d,i}^s / \tau)}{\sum_{j=1}^B \exp((\mathbf{F}_{d,i}^u)^\top \mathbf{F}_{d,j}^s / \tau)}. \quad (13)$$

Here, τ denotes the temperature parameter. In the implementation, the inverse temperature can be parameterized as a learnable logit scale. The satellite-to-UAV term $\mathcal{L}_d^{s \rightarrow u}$ is defined symmetrically. The structural contrastive loss $\mathcal{L}_s$ follows the same formulation, but is computed using the structural embeddings $\mathbf{F}_{s,i}^u$ and $\mathbf{F}_{s,i}^s$.

3.6.2 Structure-guided manifold regularization ($\mathcal{L}_r$)

To encourage the visual representation to respect the underlying structural topology, we introduce a structure-guided Laplacian regularization term $\mathcal{L}_r$.

For each matched UAV-satellite pair, we first construct a structural prototype and a discriminative prototype:

$$\mathbf{s}_i = \frac{1}{2} (\mathbf{F}_{s,i}^u + \mathbf{F}_{s,i}^s), \quad \mathbf{z}_i = \frac{1}{2} (\mathbf{F}_{d,i}^u + \mathbf{F}_{d,i}^s). \quad (14)$$

Following the implementation, both prototypes are ℓ_2 -normalized before graph construction:

$$\tilde{\mathbf{s}}_i = \frac{\mathbf{s}_i}{\|\mathbf{s}_i\|_2}, \quad \tilde{\mathbf{z}}_i = \frac{\mathbf{z}_i}{\|\mathbf{z}_i\|_2}. \quad (15)$$

We then build a structure-induced affinity graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{W})$ over the samples in the mini-batch. The adjacency matrix $\mathbf{W} \in \mathbb{R}^{B \times B}$ is computed using a symmetric Gaussian kernel in the normalized structural feature space:

$$w_{ij} = \exp \left(-\frac{\|\tilde{\mathbf{s}}_i - \tilde{\mathbf{s}}_j\|_2^2}{2\sigma^2} \right). \quad (16)$$

In practice, self-loops are removed by setting $w_{ii} = 0$, since they do not contribute to the pairwise Dirichlet energy. The degree matrix is defined as:

$$\mathbf{D}_{ii} = \sum_{j=1}^B w_{ij}, \quad (17)$$

and the combinatorial graph Laplacian is $\mathbf{L} = \mathbf{D} - \mathbf{W}$.

Let $\tilde{\mathbf{Z}} = [\tilde{\mathbf{z}}_1, \tilde{\mathbf{z}}_2, \dots, \tilde{\mathbf{z}}_B]^\top \in \mathbb{R}^{B \times d}$ denote the normalized discriminative prototypes of the mini-batch. We enforce $\tilde{\mathbf{Z}}$ to be smooth with respect to the structure-induced graph by minimizing the normalized Dirichlet energy:

$$\mathcal{L}_r = \frac{1}{B^2} \text{tr}(\tilde{\mathbf{Z}}^\top \mathbf{L} \tilde{\mathbf{Z}}) = \frac{1}{2B^2} \sum_{i,j=1}^B w_{ij} \|\tilde{\mathbf{z}}_i - \tilde{\mathbf{z}}_j\|_2^2. \quad (18)$$

This objective encourages samples with similar structural layouts to remain close in the discriminative feature space.

Therefore, the discriminative manifold is regularized by the more stable structural topology rather than being driven solely by appearance-level cues.

Definition 1 (Graph Fourier Transform on the Structure-Induced Graph). Given the structure-induced affinity matrix $\mathbf{W}$, the combinatorial graph Laplacian is defined as:

$$\mathbf{L} = \mathbf{D} - \mathbf{W}. \quad (19)$$

Since $\mathbf{W}$ is symmetric and non-negative, $\mathbf{L}$ is symmetric positive semi-definite and admits the eigendecomposition $\mathbf{L} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_B)$ contains the graph eigenvalues satisfying:

$$0 = \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_B. \quad (20)$$

The Graph Fourier Transform of the discriminative graph signal $\tilde{\mathbf{Z}}$ is defined as:

$$\tilde{\mathbf{Z}} = \mathbf{U}^\top \tilde{\mathbf{Z}}, \quad (21)$$

where each row of $\tilde{\mathbf{Z}}$ represents the spectral coefficient associated with a graph frequency.

Proposition 1 (Spectral Smoothing Induced by $\mathcal{L}_r$). When the structure-induced graph is fixed, a gradient-descent step induced by $\mathcal{L}_r$ with respect to the normalized discriminative graph signal $\tilde{\mathbf{Z}}$ acts as a graph spectral low-pass smoothing operator.

Proof. Since $\mathbf{L}$ is symmetric, the gradient of $\mathcal{L}_r$ with respect to $\tilde{\mathbf{Z}}$ is:

$$\nabla_{\tilde{\mathbf{Z}}} \mathcal{L}_r = \frac{2}{B^2} \mathbf{L} \tilde{\mathbf{Z}}. \quad (22)$$

If we analyze the effect of this regularization term alone, a gradient descent update with learning rate η gives:

$$\tilde{\mathbf{Z}}^{(t+1)} = \tilde{\mathbf{Z}}^{(t)} - \eta \frac{2}{B^2} \mathbf{L} \tilde{\mathbf{Z}}^{(t)}. \quad (23)$$

Projecting the update onto the graph spectral basis yields:

$$\begin{aligned} \tilde{\mathbf{Z}}^{(t+1)} &= \mathbf{U}^\top \tilde{\mathbf{Z}}^{(t+1)} \\ &= \mathbf{U}^\top \left(\tilde{\mathbf{Z}}^{(t)} - \eta \frac{2}{B^2} \mathbf{L} \tilde{\mathbf{Z}}^{(t)} \right) \\ &= \left(\mathbf{I} - \frac{2\eta}{B^2} \mathbf{\Lambda} \right) \tilde{\mathbf{Z}}^{(t)}. \end{aligned} \quad (24)$$

Therefore, the k -th spectral component evolves as:

$$\hat{\mathbf{z}}_k^{(t+1)} = \left(1 - \frac{2\eta}{B^2} \lambda_k \right) \hat{\mathbf{z}}_k^{(t)}. \quad (25)$$

Under the condition:

$$0 < \eta < \frac{B^2}{2\lambda_{\max}}, \quad (26)$$

the contraction factor is non-negative and decreases monotonically with λ_k . Thus, spectral components associated with larger eigenvalues, namely high graph frequencies, are attenuated more strongly, while low-frequency components are better preserved. This demonstrates that $\mathcal{L}_r$ induces a low-pass smoothing effect over the structure-induced graph. $\square$

Theorem 1 (Robustness Interpretation under Temporal Appearance Noise). Let the observed discriminative prototype signal be:

$$\tilde{\mathbf{Z}}_o = \tilde{\mathbf{Z}}_s + \mathbf{E}_t, \quad (27)$$

where $\tilde{\mathbf{Z}}_s$ denotes the structure-consistent representation and $\mathbf{E}_t$ denotes temporal appearance perturbations. If $\tilde{\mathbf{Z}}_s$ is mainly concentrated on low graph frequencies while $\mathbf{E}_t$ is mainly concentrated on high graph frequencies, then the Laplacian regularization suppresses temporal perturbations more strongly than the structure-consistent signal.

Proof. To analyze the denoising behavior of the regularization, consider the following Tikhonov-type objective:

$$\min_{\tilde{\mathbf{Z}}} J(\tilde{\mathbf{Z}}) = \|\tilde{\mathbf{Z}} - \tilde{\mathbf{Z}}_o\|_F^2 + \beta \text{tr}(\tilde{\mathbf{Z}}^\top \mathbf{L} \tilde{\mathbf{Z}}), \quad (28)$$

where $\beta > 0$ controls the strength of the Laplacian regularization. The normalization factor in Eq. (18) can be absorbed into β and therefore does not change the spectral form of the solution.

Setting the derivative of Eq. (28) with respect to $\tilde{\mathbf{Z}}$ to zero gives:

$$2(\tilde{\mathbf{Z}} - \tilde{\mathbf{Z}}_o) + 2\beta \mathbf{L} \tilde{\mathbf{Z}} = 0. \quad (29)$$

Therefore,

$$(\mathbf{I} + \beta \mathbf{L}) \tilde{\mathbf{Z}}^* = \tilde{\mathbf{Z}}_o, \quad (30)$$

and the closed-form solution is:

$$\tilde{\mathbf{Z}}^* = (\mathbf{I} + \beta \mathbf{L})^{-1} \tilde{\mathbf{Z}}_o. \quad (31)$$

Using the eigendecomposition: $\mathbf{L} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, the solution in the graph spectral domain becomes:

$$\tilde{\mathbf{Z}}^* = (\mathbf{I} + \beta \mathbf{\Lambda})^{-1} \tilde{\mathbf{Z}}_o. \quad (32)$$

For the k -th graph frequency, we have:

$$\hat{\mathbf{z}}_k^* = h(\lambda_k) (\hat{\mathbf{z}}_{s,k} + \hat{\mathbf{e}}_{t,k}), \quad h(\lambda_k) = \frac{1}{1 + \beta \lambda_k}. \quad (33)$$

Since $h(\lambda_k)$ is a monotonically decreasing function of λ_k , high-frequency components are attenuated more strongly than low-frequency components.

More specifically, for low-frequency structural components satisfying $\lambda_k \leq \lambda_l$, the response is lower-bounded by:

$$h(\lambda_k) \geq \frac{1}{1 + \beta \lambda_l}, \quad (34)$$

whereas for high-frequency temporal perturbations satisfying $\lambda_k \geq \lambda_h$, the response is upper-bounded by:

$$h(\lambda_k) \leq \frac{1}{1 + \beta \lambda_h}, \quad \lambda_h > \lambda_l. \quad (35)$$

Therefore, when temporal appearance shifts mainly violate the structure-induced affinity and appear as high-frequency graph components, they are suppressed more significantly than the low-frequency structure-consistent signal. This provides a spectral explanation for why the proposed manifold regularization improves robustness under temporal appearance variations. $\square$

4 Experiments

4.1. Datasets

To comprehensively evaluate the performance of STAR under diverse conditions, ranging from viewpoint changes to extreme environmental shifts, we conduct experiments on three benchmarks: the standard University-1652 (Zheng et al., 2020), the altitude-varying SUES-200 (Zhu et al., 2023a), and our newly constructed robustness benchmark, UniT-1652. These datasets cover two primary tasks, namely UAV-to-satellite retrieval for localization and satellite-to-UAV retrieval for navigation.

University-1652: University-1652 is a dataset specifically designed for UAV-based visual geo-localization. It contains 1,652 university buildings captured from varying viewpoints and orientations.

Table 1. Comparisons with the state-of-the-art methods on the University-1652 dataset across both UAV-view target localization (UAV→Satellite) and UAV navigation (Satellite→UAV). Bold and underlined values indicate the best and second-best results.

Method	UAV→Satellite		Satellite→UAV	
	Recall@1	AP	Recall@1	AP
LCM (Ding et al., 2020)	66.65	70.82	79.89	65.38
RK-Net (Lin et al., 2022)	68.10	71.53	80.96	69.35
MCCG (Shen et al., 2023)	89.64	91.32	94.30	89.39
ViT (Dosovitskiy et al., 2020)	74.09	77.82	83.31	72.27
LPN (T. Wang et al., 2021)	75.93	79.14	86.45	74.79
AEN (w. LPN) (Liu et al., 2024)	77.40	80.27	90.30	76.01
CA-HRS (Lu et al., 2022)	81.00	83.64	88.16	77.46
Vision (Dai et al., 2023)	82.22	84.78	87.59	81.49
OSNet (Song et al., 2023)	83.26	85.84	90.30	82.71
PCL (Tian et al., 2021)	83.27	87.32	91.78	82.18
Swin-B (Liu et al., 2021)	84.15	86.62	90.30	83.55
FSRA (Dai et al., 2021)	85.50	87.53	89.73	84.94
SDPL (S) (Chen et al., 2024)	85.57	87.61	88.73	84.75
Coarse (X. Wang et al., 2025)	85.95	90.33	94.01	82.66
SwinV2-B (Liu et al., 2022)	86.99	89.02	91.16	85.77
MFJRLN (Ge et al., 2024b)	87.61	89.57	91.07	86.72
TransFG (Zhao et al., 2024)	87.92	89.99	93.37	87.94
GeoFormer (Q. Li et al., 2024)	88.16	90.03	91.87	87.92
SDPL (B) (Chen et al., 2024)	90.16	91.64	93.58	89.45
Sample4Geo (Deuser et al., 2023)	92.65	93.81	95.14	91.39
MEAN (Chen et al., 2025)	93.55	94.53	96.01	92.08
Cities(s) (Huang et al., 2024)	93.76	94.78	<u>96.15</u>	<u>92.94</u>
Cities(b) (Huang et al., 2024)	97.43	<u>95.01</u>	96.01	92.57
STAR	<u>94.35</u>	95.30	96.43	94.13

SUES-200: SUES-200 is characterized by its inclusion of UAV images captured at multiple altitudes, with a particular focus on assessing the effect of flight height on model performance. The dataset was collected across 200 unique scenes in Shanghai. For each scene, there are 50 UAV-view images at a resolution of 1080×1080, along with one corresponding satellite-view image at a fixed altitude and resolution of 512×512. Following the experimental setup in (Xia et al., 2024), we split the dataset into a training set with 120 scenes and a test set with the remaining 80 scenes.

UniT-1652: Existing UAV geo-localization benchmarks contain limited temporal or seasonal variations, which hampers the evaluation of robustness under real-world environmental changes. To address this issue, we introduce UniT-1652, a synthetic robustness extension of University-1652 that simulates severe winter conditions. We adopt a stochastic weather injection pipeline based on Albumentations (Buslaev et al., 2020) to generate appearance shifts while preserving underlying scene structure. Concretely, UniT-1652 is stratified into three subsets according to environmental severity: (i) 50% of images remain unaltered to preserve basic structural discriminability; (ii) 25% are transformed to simulate early-

winter conditions by adding light snow with point density in [0.1, 0.3] and a brightness coefficient 1.5; and (iii) the remaining 25% represent extreme weather by combining heavy snow and fog (fog coefficient in [0.4, 0.6] with $\alpha = 0.2$), together with stronger brightness/contrast perturbations (brightness limit in [0.2, 0.4] and contrast limit in $[-0.2, 0.0]$). These augmentations are applied to both UAV and satellite views, yielding a challenging benchmark for assessing invariance to texture and illumination changes induced by seasonal shifts.

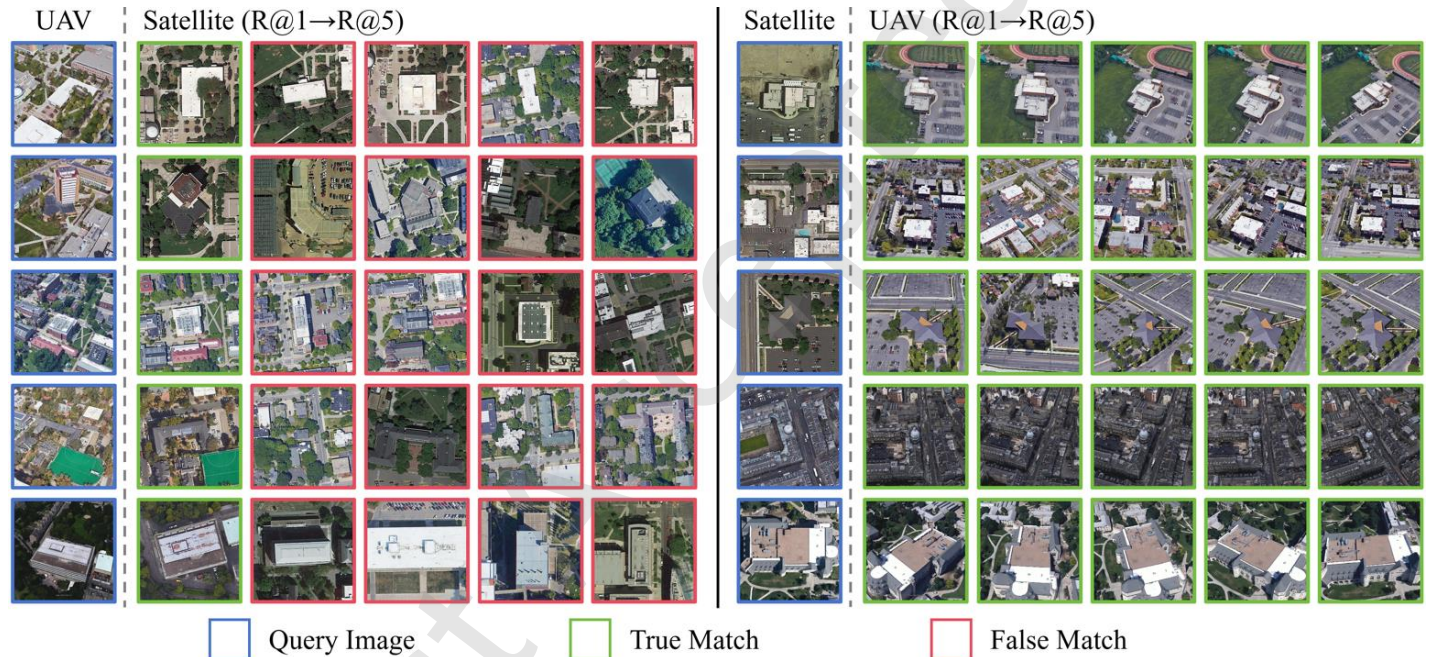
4.2. Evaluation metrics

To quantitatively evaluate STAR, we follow the common practice in UAV-based geo-localization and report Recall@K and Average Precision (AP) (Huang et al., 2024). Recall@K measures whether the correct cross-view match can be retrieved within the top-ranked candidates, while AP further evaluates the ranking quality of the retrieved list.

Recall@K: Recall@K measures the percentage of queries for which the correct match is found among the top- K retrieved candidates. Let N denote the total number of queries, and let $r_i \in \{0,1\}$ be an indicator variable denoting whether the ground-truth match for the i -th query is within the top- K results.

Table 2. Comparison with the state-of-the-art models on the SUES-200 dataset.

Method	Recall@1 (UAV→Sat)				AP (UAV→Sat)				Recall@1 (Sat→UAV)				AP (Sat→UAV)			
	150m	200m	250m	300m	150m	200m	250m	300m	150m	200m	250m	300m	150m	200m	250m	300m
LCM (Ding et al., 2020)	43.42	49.42	54.47	60.43	49.65	55.91	60.31	65.78	57.50	68.75	72.50	75.00	38.11	49.19	47.94	59.36
SUES (Zhu et al., 2023a)	55.65	66.78	72.00	74.05	61.92	71.55	76.43	78.26	75.00	85.00	86.25	88.75	55.46	66.05	69.94	74.46
LPN (T. Wang et al., 2021)	61.58	70.85	80.38	81.47	67.23	75.96	83.80	84.53	83.75	88.75	92.50	92.50	66.78	75.01	81.34	85.72
FSRA (Dai et al., 2021)	68.25	83.00	90.68	91.95	73.45	85.99	92.27	93.46	83.75	90.00	93.75	95.00	76.67	85.34	90.17	92.03
MJRLIFS (Ge et al., 2024a)	77.57	89.50	92.58	97.40	81.30	91.40	94.21	97.92	<u>93.75</u>	<u>97.50</u>	<u>97.50</u>	100.0	79.49	90.52	96.03	<u>97.66</u>
MCCG (Shen et al., 2023)	82.22	89.38	93.82	95.07	85.47	91.41	95.04	96.20	<u>93.75</u>	93.75	96.25	<u>98.75</u>	<u>89.72</u>	92.21	<u>96.14</u>	96.64
MBF (Zhu et al., 2023b)	<u>85.62</u>	87.43	90.65	92.12	<u>88.21</u>	90.02	92.53	93.63	88.75	91.25	93.75	96.25	84.74	89.95	90.65	91.60
Coarse (X. Wang et al., 2025)	76.90	87.88	92.98	95.10	84.95	92.60	95.66	96.92	87.50	92.50	96.25	<u>98.75</u>	74.81	87.15	91.20	94.52
Cities(s) (Huang et al., 2024)	77.00	80.80	87.98	92.95	81.22	84.30	90.04	94.27	88.75	92.50	<u>97.50</u>	97.50	76.18	83.09	89.45	92.52
SDPL (Chen et al., 2024)	82.95	<u>92.73</u>	<u>96.05</u>	<u>97.83</u>	85.82	<u>94.07</u>	<u>96.69</u>	<u>98.05</u>	<u>93.75</u>	96.25	<u>97.50</u>	96.25	83.75	<u>92.42</u>	95.65	96.17
STAR	91.33	97.13	98.88	99.65	93.11	97.67	99.09	99.72	98.75	98.75	98.75	97.50	91.86	96.80	98.12	97.84

**Fig. 3.** Qualitative retrieval results of STAR. Left: UAV-view target localization; Right: UAV navigation.

Recall@K is then computed as:

$$\text{Recall@K} = \frac{1}{N} \sum_{i=1}^N r_i \quad (36)$$

Average Precision: Average Precision (AP) evaluates the ranking quality of the retrieved list by summarizing the precision values at the ranks where relevant items appear.

4.3. Implementation details

STAR is implemented in PyTorch. All experiments are conducted on four NVIDIA A40 GPUs. We use DataParallel for multi-GPU training across four GPUs. We adopt TinyViT-21M (K. Wu et al., 2022) as the unified visual backbone for both UAV-view and satellite-view images. The backbone is initialized with ImageNet-pretrained weights. All input images are resized to 512×512 and normalized using the default mean and standard deviation associated with the pretrained backbone. During training, we apply the same random horizontal flip to paired UAV and satellite images with a probability of 0.5, so that the cross-view

correspondence is preserved after augmentation. For structural parsing, we use a semantic encoder pretrained on LoveDA (Hwang et al., 2024; J. Wang et al., 2021), which remains frozen during STAR training. During training, the semantic parsing branch provides structural responses for structure-aware supervision and structure-guided manifold regularization. The final discriminative retrieval feature and structural feature are both projected into a 576-dimensional embedding space. This shared dimensionality allows STAR to impose structure-aware constraints directly on the retrieval manifold without introducing an additional modality-specific matching space. Before similarity computation, the retrieval descriptors are ℓ_2 -normalized. We train STAR using the AdamW optimizer with an initial learning rate of 1×10^{-3} . Mixed-precision training is enabled with automatic mixed precision and gradient scaling, where the initial scale is set to 2^{10} . To stabilize optimization, gradient values are clipped with a threshold of

100. The learning rate is updated after each mini-batch iteration rather than only at epoch boundaries. Specifically, we adopt a cosine learning-rate schedule with a linear warm-up stage of 0.1 epochs. After warm-up, the learning rate is smoothly decayed according to the cosine schedule over the remaining training iterations. Therefore, the scheduling horizon is determined by the total number of mini-batch iterations for each training setting. For example, under the 5-epoch setting used on University-1652 and UniT-1652, one epoch accounts for 20% of the total training iterations, whereas under the 2-epoch setting used on SUES-200, one epoch accounts for 50% of the total training iterations. For University-1652 and UniT-1652, STAR is trained for 5 epochs with a mini-batch size of 128 paired samples. For SUES-200, STAR is trained separately for each altitude setting. The mini-batch size of SUES-200 is set to 24 paired samples and the model is trained for 2 epochs for each altitude setting.

4.4. Evaluation results

4.4.1 Comparison with state-of-the-art methods

To evaluate the effectiveness of STAR, we compare it with recent state-of-the-art methods on three benchmarks, including University-1652, SUES-200, and UniT-1652. These benchmarks cover both UAV-view target localization, i.e., UAV → Satellite retrieval, and UAV navigation, i.e., Satellite → UAV retrieval. The quantitative results are reported in Tables 1, 2, and 3.

Table 3. Comparisons with the state-of-the-art methods on the UniT-1652 dataset.

Method	UAV→Satellite		Satellite→UAV	
	R@1	AP	R@1	AP
ResNeXt-101 (Xie et al., 2017)	37.17	42.63	65.91	37.96
CaiT (Touvron et al., 2021)	66.26	70.69	87.59	68.09
VOLO (Yuan et al., 2022)	72.81	76.81	86.45	72.93
ConvFormer (Yu et al., 2023)	82.41	85.19	93.30	82.51
Sample4Geo (Deuser et al., 2023)	<u>87.77</u>	<u>89.84</u>	<u>95.01</u>	<u>87.34</u>
STAR	92.71	94.01	96.29	92.86

On University-1652, STAR achieves competitive performance compared with existing methods. For UAV→Satellite retrieval, STAR obtains the highest AP of 95.30, improving over the strongest competing AP result, i.e., Cities(b), by 0.29 percentage points. Although Cities(b) reports a higher Recall@1, STAR ranks second on this metric and achieves better AP, suggesting that its advantage is reflected more clearly in the overall retrieval ranking. For Satellite→UAV retrieval, STAR obtains the best results on both Recall@1 and AP. Compared with the strongest previously reported results, STAR improves Recall@1 by 0.28 percentage points and AP by 1.19 percentage points. The improvement in AP indicates that the learned structure-aware representation helps produce more stable ranking results for UAV navigation. On SUES-200, STAR shows clear gains under altitude-induced viewpoint and scale variations.

For UAV→Satellite retrieval, STAR achieves the best results across all four altitudes. Compared with the strongest reported baseline at each altitude, STAR improves Recall@1 by 5.71, 4.40, 2.83, and 1.82 percentage points from 150m to 300m, respectively. The corresponding AP improvements are 4.90, 3.60, 2.40, and 1.67 percentage points. These gains are larger at lower altitudes, where UAV images usually contain stronger perspective distortion and larger appearance differences from satellite images. This observation suggests that structural cues are more useful when direct appearance correspondence becomes less reliable. For Satellite→UAV retrieval on SUES-200, STAR also performs well across different altitudes. At 150m, where the cross-view discrepancy is relatively large, STAR improves Recall@1 over the best reported baseline by 5.00 percentage points and improves AP by 2.14 percentage points. As the altitude increases, the performance gap becomes smaller, which is reasonable because the UAV view gradually becomes closer to the satellite view. At 300m, STAR is slightly lower than MJRLIFS in Recall@1 but remains competitive, and it achieves the best AP with a marginal improvement of 0.18 percentage points over MJRLIFS. These results indicate that STAR is particularly effective in low-altitude scenarios, where viewpoint changes make appearance-based matching more difficult. On UniT-1652, which is designed to evaluate robustness under temporal domain shifts, STAR achieves the best performance in both retrieval directions. Compared with Sample4Geo, the strongest baseline in this benchmark, STAR improves UAV→Satellite Recall@1 and AP by 4.94 and 4.17 percentage points, respectively. For Satellite→UAV retrieval, STAR improves Recall@1 by 1.28 percentage points and AP by 5.52 percentage points. The larger AP improvement suggests that STAR better preserves the retrieval ranking under temporal appearance changes. A cross-benchmark comparison further supports this observation. From University-1652 to UniT-1652, Sample4Geo drops by 4.88 percentage points in UAV→Satellite Recall@1 and by 4.05 percentage points in Satellite→UAV AP, whereas STAR drops by 1.64 and 1.27 percentage points on the same metrics. This comparison shows that the proposed structure-aware representation provides more stable cues when the visual appearance of the same location changes over time. The performance gains are evident under challenging conditions involving large viewpoint gaps, altitude variations, and temporal shifts.

Overall, STAR achieves strong and consistent performance across the three benchmarks, ranking first or second on most reported metrics under diverse viewpoint and temporal-shift conditions. Taken together, these observations on three benchmarks suggest that incorporating structural guidance into representation learning helps construct a more reliable UAV-satellite retrieval space, improving both retrieval accuracy and ranking stability in challenging UAV-based geo-localization scenarios.

Table 4. The efficiency and performance of models on the University-1652 dataset.

Method	#Para(M)	GFLOPs	UAV→Satellite		Satellite→UAV	
			Recall@1	AP	Recall@1	AP
LPN (T. Wang et al., 2021)	62.4	36.78	75.93	79.14	86.45	74.79
SDPL (B) (Chen et al., 2024)	42.6	69.71	90.16	91.64	93.58	89.45
Sample4Geo (Deuser et al., 2023)	88.6	90.24	92.65	93.81	95.14	91.39
MEAN (Chen et al., 2025)	36.5	26.18	93.55	94.53	96.01	92.08
Cities(s) (Huang et al., 2024)	<u>22.7</u>	<u>24.05</u>	93.76	94.78	<u>96.15</u>	<u>92.94</u>
Cities(b) (Huang et al., 2024)	90.5	91.63	97.43	<u>95.01</u>	96.01	92.57
STAR	21.2	21.31	<u>94.35</u>	95.30	96.43	94.13

4.4.2 Efficiency comparison

Efficiency is a critical requirement for real-world UAV-based geo-localization systems, where onboard computational resources are often limited. As reported in Table 4, STAR achieves the best overall performance while maintaining the smallest parameter count (21.2M) and the lowest computational cost (21.31 GFLOPs) among all compared methods. Compared to the best baseline, Cities(b) (Huang et al., 2024), STAR reduces the number of parameters and computational complexity by more than 75%, while still surpassing it in three out of four evaluation metrics. Moreover, STAR consistently outperforms the lightweight variant Cities(s) across all metrics, demonstrating a substantially better balance between accuracy and efficiency. To further evaluate the training-phase cost introduced by the structural guidance, we report the training efficiency of STAR in Table 5. The reported GPU memory denotes the total memory consumption across four NVIDIA A40 GPUs, and the training time is measured by the average time required for one training epoch. These results indicate that although STAR introduces additional training-time structural components, the overall training cost remains manageable under the reported experimental settings.

Table 5. Training-phase efficiency of STAR.

Dataset	GPU Memory	Epochs	Time/epoch
University-1652	~120 GiB	5	~5 min
SUES-200	~25 GiB	2	~3 min

4.4.3 Qualitative results

Fig. 3 presents qualitative retrieval results on the University-1652 dataset. In each example, STAR retrieves the correct match within the top-5 candidates, showing that the learned representation can identify the corresponding satellite image under cross-view appearance changes. The retrieved candidates also indicate that STAR maintains reasonable ranking quality across different scene types, which is consistent with the quantitative results reported above. These examples provide an intuitive illustration of the model's ability to handle

viewpoint differences in UAV-based geo-localization. To further illustrate the behavior of the Semantic Encoder, we visualize its semantic segmentation outputs on the LoveDA dataset (J. Wang et al., 2021). As shown in Fig. 4, we present the original image, the predicted semantic map, and the overlay between the semantic predictions and the input image. The Semantic Encoder can identify major structural elements, such as roads, buildings, and vegetation, and produces generally coherent semantic regions. These visualizations suggest that the encoder is able to provide useful high-level structural cues, which can support the structure-aware representation learning process in STAR.

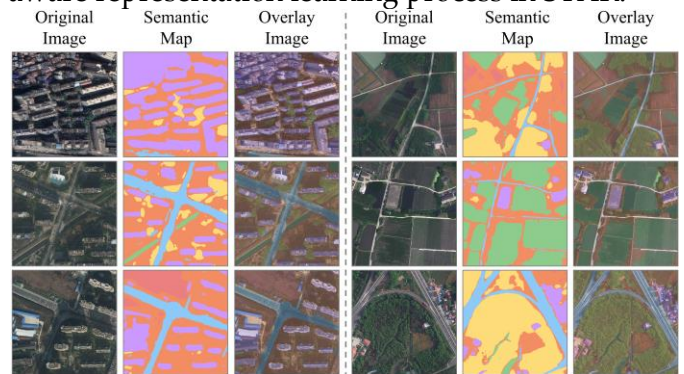


Fig. 4. Qualitative results of the Semantic Encoder. Each triplet shows the original image (left), the predicted semantic map (middle), and the overlay of the semantic map with the input image (right). The results demonstrate that the Semantic Encoder effectively captures structural cues (e.g., roads, buildings, vegetation), forming a robust intermediate representation for structure-aware geo-localization.

4.5 Ablation studies

4.5.1 Impact of each module on temporal robustness

To investigate how different components contribute to temporal robustness, we conduct an ablation study on UniT-1652 by comparing three model variants: 1) STAR (w/o SS), which removes the Structural Stream and retains only the TinyViT appearance backbone; 2) STAR (w/o SD), which

preserves the Structural Stream but replaces the graph-based Structure-Aware Decoder with a convolutional decoder; and 3) the full STAR model. The quantitative results are reported in Table 6, and representative qualitative examples are shown in Fig. 5. As shown in Table 6, the introduction of the Structural Stream leads to consistent improvements in both retrieval directions. Compared with STAR (w/o SS), the full STAR improves Recall@1/AP by 0.58/0.47 percentage points for UAV→Satellite retrieval and by 0.43/0.62 percentage points for Satellite→UAV retrieval. These results indicate that structural semantics provide complementary information to appearance features, which helps reduce the influence of temporal appearance variations and improves the stability of cross-temporal matching. The contribution of the Structure-Aware Decoder can be observed by comparing STAR (w/o SD) with the full model. With the graph-based decoder, STAR obtains higher Recall@1 and AP in UAV→Satellite retrieval, and further improves AP in Satellite→UAV retrieval while preserving the best Recall@1. This result suggests that the Structure-Aware Decoder facilitates the use of structural information by modeling relations among semantic regions.

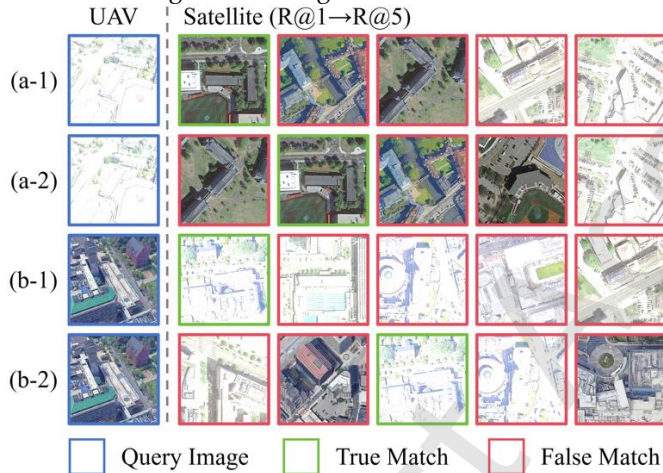


Fig. 5. Qualitative comparison under temporal shifts on UniT-1652. Each query UAV image is matched against a satellite gallery, and the top-5 retrieved satellite images are shown from left to right. Green and red boxes denote true and false matches, respectively. In both examples, the full STAR model, shown in (a-1) and (b-1), retrieves the correct match at Rank-1, whereas the TinyViT-only variant, shown in (a-2) and (b-2), places the true match at a lower rank. These examples illustrate that structure-aware representation helps preserve stable geographic layout cues and improves reliability under temporal changes.

The qualitative results in Fig. 5 further illustrate the effect of structure-aware representation learning under temporal shifts. When the UAV query and the satellite gallery exhibit visible seasonal or illumination differences, the TinyViT-only variant is more prone to retrieving candidates with similar local appearance but incorrect geographic correspondence. In contrast, STAR retrieves the correct satellite image at Rank-1 in both examples. This observation is consistent with the quantitative results and suggests that

structural guidance helps preserve stable geographic layout information during retrieval. Overall, the ablation results show that both the Structural Stream and the Structure-Aware Decoder contribute to the temporal robustness of STAR. The Structural Stream introduces semantic structural cues beyond appearance, while the Structure-Aware Decoder further organizes these cues through relational modeling. Together, these components support the learning of a structure-aware retrieval representation that is more robust to temporal domain shifts.

Table 6. Ablation study of STAR on UniT-1652 for temporal robustness. SS and SD denote the Structural Stream and the Structure-Aware Decoder, respectively. STAR (w/o SS) removes the Structural Stream and only retains the TinyViT appearance backbone, while STAR (w/o SD) replaces the graph-based Structure-Aware Decoder with a convolutional decoder.

Method	UAV→Satellite		Satellite→UAV	
	Recall@1	AP	Recall@1	AP
STAR (w/o SS)	92.13	93.54	95.86	92.24
STAR (w/o SD)	92.28	93.68	96.29	92.72
STAR	92.71	94.01	96.29	92.86

4.5.2 Module contribution under structural masking

To further examine how STAR behaves when local structural cues are partially perturbed, we conduct a controlled structural masking experiment on the University-1652 dataset during both training and evaluation. In this setting, local scene regions related to structural information are masked to simulate partial loss or disturbance of local structural cues. Under this setting, we analyze the contribution of two key components in STAR, namely the Structural Stream and the Structure-Aware Decoder. The quantitative results are reported in Table 7. As shown in Table 7, removing the Structural Stream leads to performance degradation in both retrieval directions. Compared with STAR (w/o SS), the full STAR achieves consistent improvements in Recall@1 and AP for both retrieval directions. The improvement is particularly evident in the Satellite→UAV direction, where the model needs to retrieve UAV-view images from a satellite-view query under local structural perturbations. This result indicates that structural cues provide complementary information to appearance features and help improve retrieval stability when parts of the scene structure are masked or altered. The contribution of the Structure-Aware Decoder is also reflected in the comparison between STAR (w/o SD) and the full model. Replacing the graph-based decoder with a convolutional decoder reduces the performance across all metrics. This suggests that the structural branch benefits not only from semantic structural supervision, but also from relational modeling among structural regions. As structural masking represents a synthetic perturbation, this experiment evaluates robustness to controlled structural perturbations rather than complex real-world structural evolution.

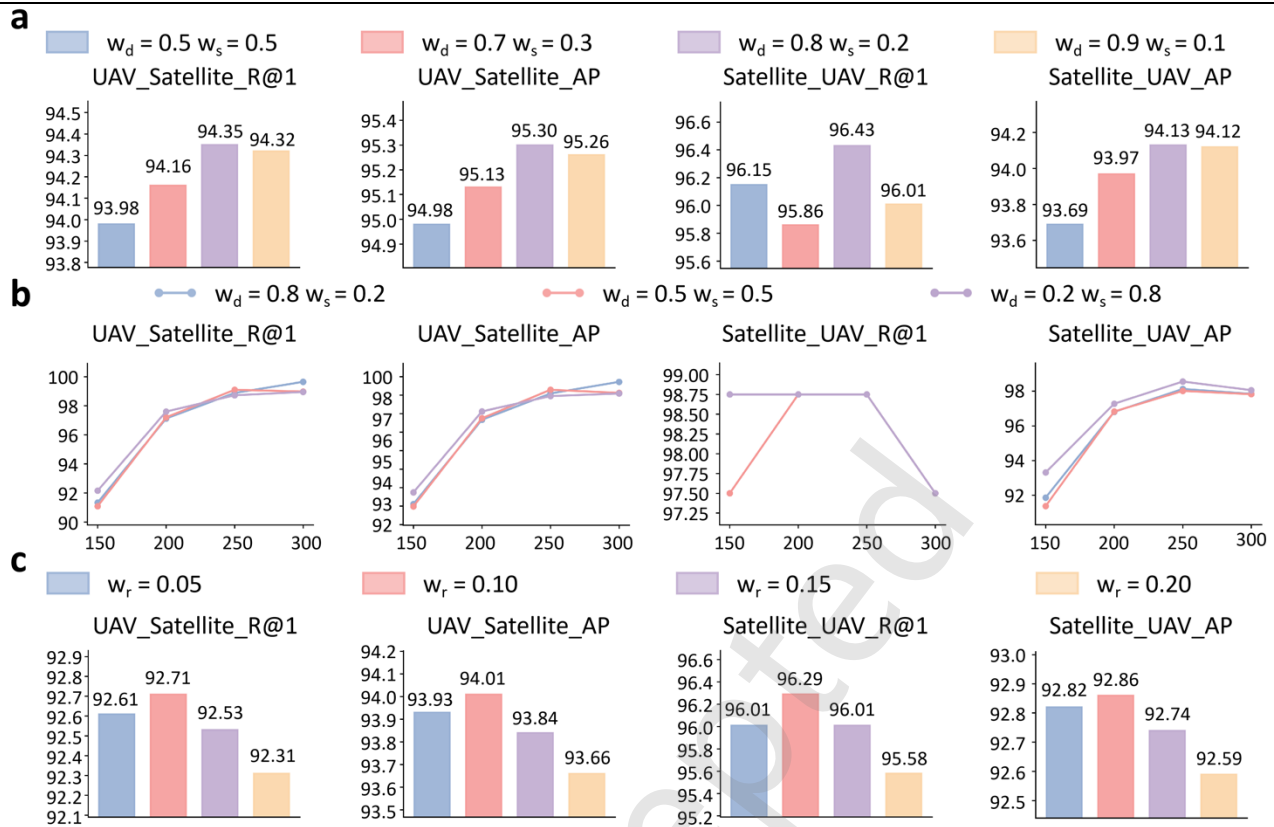


Fig. 6. Sensitivity analysis of loss weights in STAR. (a) Influence of the discriminative retrieval loss weight w_d and the structural learning loss weight w_s on University-1652, with w_r fixed to 0.10. (b) Influence of w_d and w_s on SUES-200 under different UAV flight altitudes. Lower altitudes introduce larger cross-view discrepancies, under which stronger structural supervision becomes more beneficial, while higher altitudes favor appearance-dominant discrimination. (c) Influence of the structure-guided manifold regularization weight w_r on UniT-1652, with $w_d = 0.80$ and $w_s = 0.20$ fixed.

Table 7. Ablation study of STAR on University-1652 under structural masking. SS and SD denote the Structural Stream and the Structure-Aware Decoder, respectively.

Method	UAV→Satellite		Satellite→UAV	
	Recall@1	AP	Recall@1	AP
STAR (w/o SS)	91.62	93.08	93.87	90.29
STAR (w/o SD)	90.53	92.13	94.72	89.99
STAR	91.88	93.31	95.15	90.91

4.5.3 Impact of structural awareness on noise robustness

The structural branch of STAR relies on semantic maps as training-time guidance. In practice, semantic predictions may contain errors, which can introduce noisy structural supervision into the optimization process. To examine the influence of semantic-map quality on geo-localization performance, we inject controlled noise into the semantic maps during training and evaluate the model under three noise levels, i.e., 10%, 30%, and 50%. We compare the full STAR model with STAR (w/o SD), where the graph-based Structure-Aware Decoder is replaced by a convolutional decoder. This comparison is used to assess the role of relational structural modeling when the semantic guidance is imperfect. As reported in Table 8, STAR consistently outperforms STAR (w/o SD) under all semantic-noise levels

and in both retrieval directions. At the 10% noise level, the full model improves the average performance across the four metrics by 3.01 percentage points over the convolutional-decoder variant. The improvements are observed in both Recall@1 and AP, indicating that the graph-based decoder helps maintain retrieval accuracy and ranking quality under mild semantic perturbations. When the noise level increases to 30%, STAR continues to achieve higher performance than STAR (w/o SD), with an average gain of 1.91 percentage points across the four metrics. This result suggests that explicitly modeling relations among semantic regions can provide more stable structural guidance than local convolutional decoding when semantic maps become less reliable. At the 50% noise level, STAR still maintains clear advantages, with particularly notable gains in UAV→Satellite Recall@1 and Satellite→UAV AP. Averaged over all metrics and all noise levels, the full STAR improves over STAR (w/o SD) by 2.81 percentage points. Overall, these results show that the Structure-Aware Decoder improves the tolerance of STAR to controlled semantic-map perturbations. By aggregating semantic information through graph-based relational modeling, STAR can better exploit structural cues under imperfect semantic maps and preserve a stable structure-aware retrieval representation for robust UAV-based geo-localization.

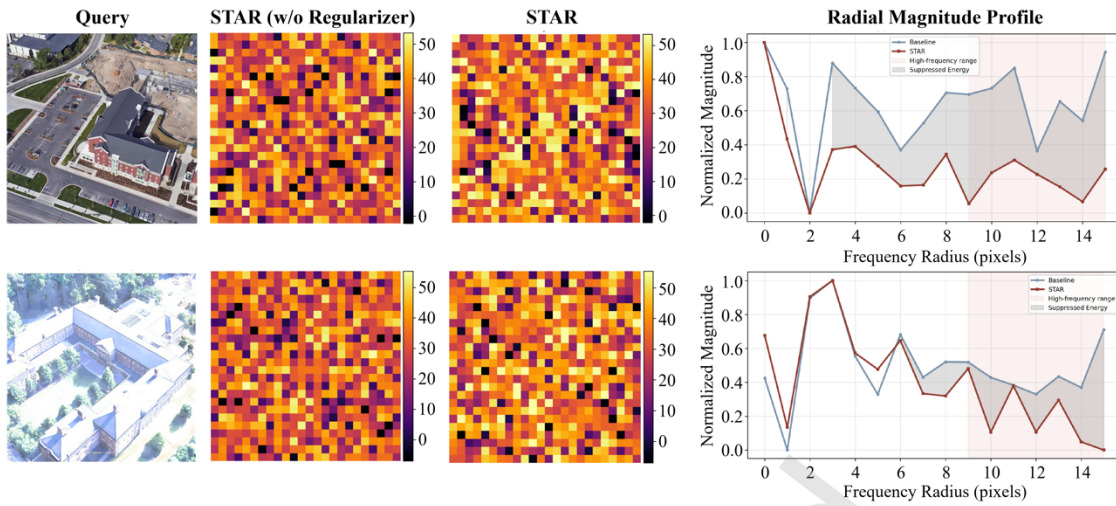


Fig. 7. Feature-domain spectral visualization on UniT-1652. For each query, the middle columns show the log-magnitude spectra of STAR without structure-guided manifold regularization and the full STAR model. The right column shows normalized radial magnitude profiles. The pale-red background marks the designated high-frequency range, and gray shading highlights regions where the unregularized normalized profile exceeds that of STAR.

Table 8. Ablation study on University-1652 under different levels of semantic-map noise. SD denotes the Structure-Aware Decoder.

Noise	Method	UAV→Satellite		Satellite→UAV	
		Recall@1	AP	Recall@1	AP
10%	STAR (w/o SD)	88.20	89.96	91.58	88.07
	STAR	91.62	92.92	93.87	91.43
30%	STAR (w/o SD)	87.48	89.37	91.58	87.47
	STAR	89.61	91.25	93.01	89.67
50%	STAR (w/o SD)	84.53	86.81	88.73	84.24
	STAR	88.37	90.17	91.44	88.40

4.5.4 Sensitivity analysis of loss weights

The overall objective of STAR is controlled by three loss weights, i.e., w_d , w_s , and w_r , which correspond to the discriminative retrieval loss, the structural learning loss, and the structure-guided manifold regularization, respectively. To analyze their influence, we first vary w_d and w_s with w_r fixed, and then examine the effect of w_r with w_d and w_s fixed. The results on University-1652, SUES-200, and UniT-1652 are summarized in Fig. 6. Fig. 6(a) shows the influence of balancing the discriminative retrieval loss and the structural learning loss on University-1652. When w_r is fixed to 0.10, the setting $w_d = 0.80$ and $w_s = 0.20$ achieves the best overall performance across the two retrieval directions. Compared with a more appearance-dominant setting, increasing the weight of the structural objective provides consistent improvements, indicating that structural supervision offers complementary guidance beyond appearance-based discrimination. However, assigning an excessively large weight to the structural objective does not further improve the results. This suggests that the discriminative retrieval loss should remain the primary objective, while the structural learning loss serves as an auxiliary constraint to enhance representation robustness. Fig. 6(b) further evaluates the effect of w_d and w_s on SUES-200 under different UAV flight altitudes. The results show that the preferred balance between appearance discrimination and

structural supervision varies with the degree of cross-view discrepancy. At lower altitudes, where UAV images exhibit stronger viewpoint and scale differences from satellite images, increasing the structural weight leads to better retrieval performance. In contrast, at higher altitudes, the UAV viewpoint becomes closer to the satellite viewpoint, and the appearance-dominant setting achieves more favorable results. This trend indicates that structural cues become more useful when direct appearance correspondence is weakened by severe viewpoint variations, whereas appearance features become more reliable when the cross-view gap is reduced. For Satellite→UAV retrieval on SUES-200, the Recall@1 values are generally close across different weight configurations. Nevertheless, AP remains sensitive to the loss balance and reflects the influence of structural supervision on ranking quality. In several altitude settings, increasing the structural weight improves AP, suggesting that structure-aware learning helps produce more reliable retrieval rankings when multiple UAV-view candidates share similar visual appearances. Fig. 6(c) analyzes the effect of the manifold regularization weight w_r on UniT-1652, with $w_d = 0.80$ and $w_s = 0.20$ fixed. The best performance is obtained when $w_r = 0.10$. A smaller value weakens the structure-induced manifold constraint, while a larger value makes the regularization less balanced with the discriminative retrieval objective. This result indicates that

the proposed manifold regularization should be applied with a moderate weight, so that structural consistency can be incorporated without compromising the discriminative capacity of the retrieval representation. Overall, the sensitivity analysis shows that STAR benefits from a balanced combination of appearance discrimination, structural supervision, and manifold regularization.

4.5.5 Effect of structure-guided manifold regularization

To further examine the representation characteristics associated with the proposed Structure-Guided Manifold Regularization, we provide an auxiliary Fourier-domain visualization of the learned discriminative feature F_d . Specifically, each 576-dimensional feature vector is reshaped into a 24×24 map for diagnostic analysis, followed by a 2D Fast Fourier Transform. We then compute a radial magnitude profile by averaging the FFT magnitude at each frequency radius, followed by normalization for relative comparison across models. Fig. 7 compares STAR without the regularization term and the full STAR model on representative samples from UniT-1652. The unregularized variant generally exhibits stronger normalized responses in the high-frequency range of the radial profiles. These high-frequency responses correspond to rapidly varying components in the reshaped feature representation. In contrast, the full STAR model presents lower normalized high-frequency responses over most of the highlighted frequency range, as indicated by the radial profiles and the gray shaded regions. This visualization provides a complementary diagnostic perspective on the behavior associated with Structure-Guided Manifold Regularization.

5 Conclusion

This paper presents STAR, a lightweight structure-aware representation learning framework for robust UAV-based geo-localization under temporal shifts and cross-view viewpoint discrepancies. The key idea of STAR is to exploit geometrically stable structural semantics, such as road layouts, building footprints, and land-cover distributions, to regularize the discriminative retrieval manifold. By incorporating structural guidance during representation learning, STAR shapes a compact structure-aware embedding space that preserves cross-view discriminability while improving robustness to appearance variations caused by seasonal changes, illumination differences, and perspective distortions. In addition, we introduce UniT-1652, a new benchmark derived from University-1652 to evaluate UAV-based geo-localization under deployment-time temporal shifts. Extensive experiments demonstrate that STAR achieves state-of-the-art performance across multiple retrieval settings, with particularly notable advantages in challenging scenarios involving severe viewpoint gaps and temporal misalignment between UAV and satellite imagery. Comprehensive ablation studies further verify the effectiveness of structural supervision and structure-guided manifold regularization, confirming that stable geographic structures provide complementary guidance beyond

appearance-based discrimination.

The current evaluation mainly focuses on synthetically introduced temporal shifts and controlled perturbations. Future research toward deployment-oriented UAV geo-localization may further benefit from longitudinal UAV-satellite benchmarks that capture real-world structural changes and long-term environmental evolution.

Author contributions

Chengyue Wang: Conceptualization, Methodology, Experiment, Writing. Bin Rao: Methodology, Experiment, Writing. Yanchen Guan: Methodology, Review. Jiaxun Zhang: Methodology, Review. Xingcheng Liu: Methodology, Review. Haicheng Liao: Methodology, Review. Zhenning Li: Conceptualization, Methodology, Writing.

Replication and data sharing

The source code and replication package are available on ETS data at DOI: 10.26599/ETSD.2026.9190088.

Acknowledgements

This work was supported by the Science and Technology Development Fund of Macau [0007/2025/RIC, 0122/2024/RIB2, 0215/2024/AGJ, 0074/2025/AMJ, 001/2024/SKL, 0002/2025/EQP], the Research Services and Knowledge Transfer Office, University of Macau [SRG2023-00037-IOTSC, MYRG-GRG2024-00284-IOTSC], the Shenzhen-Hong Kong-Macau Science and Technology Program Category C [SGDX20230821095159012], the Science and Technology Planning Project of Guangdong [2025A0505010016], National Natural Science Foundation of China [52572354], the State Key Lab of Intelligent Transportation System [2024-B001], and the Jiangsu Provincial Science and Technology Program [BZ2024055].

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT (OpenAI) to assist with language editing, including improving grammar, clarity, and readability. After utilizing this tool/service, the author(s) thoroughly reviewed and edited the content as necessary and take full responsibility for the final content of the published article.

Declaration of competing interest

The author Zhenning Li is a Youth Editorial Board member of this journal.

References

- Alehdaghi, M., Josi, A., Cruz, R.M.O., Shamsolmoali, P., Granger, E., 2025. Adaptive generation of privileged intermediate information for visible-infrared person re-identification. *IEEE Transactions on Information Forensics and Security*, 20, 3400–3413.
- Buslaev, A., Iglovikov, V.I., Khvedchenya, E., Parinov, A.,

- Druzhinin, M., Kalinin, A.A., 2020. Albumentations: Fast and flexible image augmentations. *Information*, 11(2), 125.
- Chen, Q., Wang, T., Yang, Z., Li, H., Lu, R., Sun, Y., Zheng, B., et al., 2024. SDPL: Shifting-dense partition learning for UAV-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(11), 11810–11824.
- Chen, Z., Yang, Z.X., Rong, H.J., 2025. Multilevel embedding and alignment network with consistency and invariance learning for cross-view geo-localization. *IEEE Transactions on Geoscience and Remote Sensing*, 63, 1–15. <https://doi.org/10.1109/TGRS.2025.3572775>
- Dai, M., Hu, J., Zhuang, J., Zheng, E., 2021. A transformer-based feature segmentation and region alignment method for UAV-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7), 4376–4389.
- Dai, M., Zheng, E., Feng, Z., Qi, L., Zhuang, J., Yang, W., 2023. Vision-based UAV self-positioning in low-altitude urban environments. *IEEE Transactions on Image Processing*, 33, 493–508.
- Deuser, F., Habel, K., Oswald, N., 2023. Sample4geo: Hard negative sampling for cross-view geo-localisation. In: *Proceedings of the IEEE/CVF international conference on computer vision*, 16847–16856.
- Ding, L., Zhou, J., Meng, L., Long, Z., 2020. A practical cross-view image matching method between UAV and satellite for UAV-based geo-localization. *Remote Sensing*, 13(1), 47.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Gao, R., Qi, Q., Mei, P., Huang, C., 2025. A joint state and fault fusion estimation scheme for mobile robot localization with energy harvesting sensors. *CHAIN*, 2(1), 57–71.
- Ge, F., Zhang, Y., Liu, Y., Wang, G., Coleman, S., Kerr, D., Wang, L., 2024a. Multibranch joint representation learning based on information fusion strategy for cross-view geo-localization. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1–16.
- Ge, F., Zhang, Y., Wang, L., Liu, W., Liu, Y., Coleman, S., Kerr, D., 2024b. Multilevel feedback joint representation learning network based on adaptive area elimination for cross-view geo-localization. *IEEE transactions on geoscience and remote sensing*, 62, 1–15.
- Gu, A., Dao, T., 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.
- Guo, Y., Choi, M., Li, K., Boussaid, F., Bennamoun, M., 2022. Soft exemplar highlighting for cross-view image-based geo-localization. *IEEE transactions on image processing*, 31, 2094–2105.
- Huang, G., Zhou, Y., Zhao, L., Gan, W., 2024. Cv-cities: Advancing cross-view geo-localization in global cities. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18, 1592–1606.
- Hwang, G., Jeong, J., Lee, S.J., 2024. SFA-net: Semantic feature adjustment network for remote sensing image segmentation. *Remote Sensing*, 16(17), 3278.
- Li, Q., Yang, X., Fan, J., Lu, R., Tang, B., Wang, S., Su, S., 2024. GeoFormer: An effective transformer-based siamese network for UAV geolocalization. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17, 9470–9491.
- Li, Z., Cui, Z., Liao, H., Ash, J., Zhang, G., Xu, C., Wang, Y., 2024. Steering the future: Redefining intelligent transportation systems with foundation models. *CHAIN*, 1(1), 46–53.
- Liao, H., Shen, H., Li, Z., Wang, C., Li, G., Bie, Y., Xu, C., 2024a. GPT-4 enhanced multimodal grounding for autonomous driving: Leveraging cross-modal attention with large language models. *Communications in Transportation Research*, 4, 100116.
- Liao, H., Li, Y., Li, Z., Wang, C., Li, G., Tian, C., Bian, Z., et al., 2024b. Less is more: Efficient brain-inspired learning for autonomous driving trajectory prediction. In: *ECAI 2024*, 4361–4368.
- Lin, J., Zheng, Z., Zhong, Z., Luo, Z., Li, S., Yang, Y., Sebe, N., 2022. Joint representation learning and keypoint detection for cross-view geo-localization. *IEEE Transactions on Image Processing*, 31, 3780–3792.
- Liu, Y., Jourabloo, A., Liu, X., 2018. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 389–398.
- Liu, T., Luo, F., Xia, Q., Ma, S., Chang, B., Sui, Z., 2019. Hierarchical encoder with auxiliary supervision for neural table-to-text generation: Learning better representation for tables. In: *Proceedings of the AAAI conference on artificial intelligence*, 6786–6793.
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., et al., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: *Proceedings of the IEEE/CVF international conference on computer vision*, 9992–10002.
- Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., et al., 2022. Swin transformer v2: Scaling up capacity and resolution. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12009–12019.
- Liu, C., Li, S., Du, C., Qiu, H., 2024. Adaptive global embedding learning: A two-stage framework for uav-view geo-localization. *IEEE Signal Processing Letters*, 31, 1239–1243.
- Lu, Z., Pu, T., Chen, T., Lin, L., 2022. Content-aware hierarchical representation selection for cross-view geo-localization. In: *Proceedings of the asian conference on computer vision*, 4211–4224.
- Shen, T., Wei, Y., Kang, L., Wan, S., Yang, Y.H., 2023. MCCG: A ConvNeXt-based multiple-classifier method for cross-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(3), 1456–1468.
- Shi, Y., Yu, X., Liu, L., Zhang, T., Li, H., 2020. Optimal feature transport for cross-view image geo-localization. In: *Proceedings of the AAAI conference on artificial intelligence*, 11990–11997.
- Song, H., Wang, Z., Lei, Y., Shi, D., Tong, X., Lei, Y., Qiu, C., 2023. Learning visual representation clusters for cross-view geo-location. *IEEE Geoscience and Remote Sensing Letters*, 20, 1–

- 5.
- Tan, M., Le, Q., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning, 6105–6114.
- Tian, X., Shao, J., Ouyang, D., Shen, H.T., 2021. UAV-satellite view synthesis for cross-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7), 4804–4815.
- Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., Jégou, H., 2021. Going deeper with image transformers. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV), 32–42.
- Vapnik, V., Vashist, A., 2009. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5–6), 544–557.
- Wang, T., Zheng, Z., Yan, C., Zhang, J., Sun, Y., Zheng, B., Yang, Y., 2021. Each part matters: Local patterns facilitate cross-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(2), 867–879.
- Wang, J., Zheng, Z., Ma, A., Lu, X., Zhong, Y., 2021. LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*.
- Wang, T., Fan, S., Liu, D., Sun, C., 2022. Transformer-guided convolutional neural network for cross-view geolocalization. *arXiv preprint arXiv:2204.09967*.
- Wang, X., Zhang, L., Fan, Z., Liu, Y., Chen, C., Deng, F., 2025. From coarse to fine: A matching and alignment framework for unsupervised cross-view geo-localization. In: Proceedings of the AAAI conference on artificial intelligence, 8024–8032.
- Wu, H., Xiao, F., Liang, C., 2022. Dual contrastive learning with anatomical auxiliary supervision for few-shot medical image segmentation. In: European conference on computer vision, 417–434.
- Wu, Z., Xia, X., Wang, R., Li, J., Yu, J., Mao, Y., Liu, T., 2022. LR-SVM+: Learning using privileged information with noisy labels. *IEEE Transactions on Multimedia*, 24, 1080–1092. <https://doi.org/10.1109/TMM.2021.3116417>
- Wu, K., Zhang, J., Peng, H., Liu, M., Xiao, B., Fu, J., Yuan, L., 2022. Tinyvit: Fast pretraining distillation for small vision transformers. In: European conference on computer vision, 68–85.
- Xia, P., Wan, Y., Zheng, Z., Zhang, Y., Deng, J., 2024. Enhancing cross-view geo-localization with domain alignment and scene consistency. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(12), 13271–13281.
- Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K., 2017. Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 5987–5995.
- Yang, X., Wang, M., Tao, D., 2017. Person re-identification with metric learning using privileged information. *IEEE Transactions on Image Processing*, 27(2), 791–805.
- You, S., Xu, C., Wang, Y., Xu, C., Tao, D., 2017. Privileged multi-label learning. In: Proceedings of the 26th international joint conference on artificial intelligence, 3336–3342.
- Yu, W., Si, C., Zhou, P., Luo, M., Zhou, Y., Feng, J., Yan, S., et al., 2023. Metaformer baselines for vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2), 896–912.
- Yuan, L., Hou, Q., Jiang, Z., Feng, J., Yan, S., 2022. Volo: Vision outlooker for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 45(5), 6575–6586.
- Yuan, J., Zhao, W., Zhou, Y., Huang, Y., 2025. Flexible sensor-driven smart vehicles: Opportunities and prospects. *CHAIN*, 2(3), 227–235.
- Zhang, X., Li, X., Sultani, W., Zhou, Y., Wshah, S., 2023. Cross-view geo-localization via learning disentangled geometric layout correspondence. In: Proceedings of the AAAI conference on artificial intelligence, 3480–3488.
- Zhang, X., Li, X., Sultani, W., Chen, C., Wshah, S., 2024. GeoDTR+: Toward generic cross-view geolocalization via geometric disentanglement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12), 10419–10433.
- Zhao, H., Ren, K., Yue, T., Zhang, C., Yuan, S., 2024. TransFG: A cross-view geo-localization of satellite and UAVs imagery pipeline using transformer-based feature aggregation and gradient guidance. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1–12.
- Zheng, Z., Wei, Y., Yang, Y., 2020. University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In: Proceedings of the 28th ACM international conference on multimedia, 1395–1403.
- Zhu, R., Yin, L., Yang, M., Wu, F., Yang, Y., Hu, W., 2023a. SUES-200: A multi-height multi-scene cross-view image benchmark across drone and satellite. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9), 4825–4839.
- Zhu, R., Yang, M., Yin, L., Wu, F., Yang, Y., 2023b. UAV's status is worth considering: A fusion representations matching method for geo-localization. *Sensors*, 23(2), 720.

Author biography



Chengyue Wang is currently pursuing a Ph.D. degree at the State Key Laboratory of Internet of Things for Smart City, University of Macau. He received his M.S. degree in civil engineering from the University of Illinois Urbana-Champaign (UIUC) in 2022. He received his B.E. degree in transportation engineering from Chang'an University in 2021. His research interests include autonomous driving and the application of deep learning to autonomous driving.



Bin Rao received the B.S. and M.S. degrees in traffic engineering from the South China University of Technology, Guangzhou, China, in 2021 and 2024. He is currently pursuing a Ph.D. degree at the State Key Laboratory of Internet of Things for Smart City, University of Macau. His research

interests include autonomous driving trajectory prediction and planning, and accident anticipation.



Yanchen Guan is currently a PhD student at the State Key Laboratory of Internet of Things for Smart City at the University of Macau. He holds a master's degree in Mobility Engineering from Politecnico di Milano (2023) and a bachelor's degree in Mechatronics Engineering from Harbin Institute of Technology (2019). His research primarily focuses on autonomous driving and intelligent transportation systems.



Jiaxun Zhang is currently pursuing a Ph.D. degree at the State Key Laboratory of Internet of Things for Smart City, University of Macau. She holds an M.S. degree in Integrated Sustainable Design from the National University of Singapore (2023) and a B.E. degree in Traffic Engineering from the South China University of Technology (2022). Her research primarily focuses on autonomous driving.



Xingcheng Liu received the B.S. degree in mathematics from Shanghai Jiao Tong University. He is currently a Ph.D. student at the State Key Laboratory of Internet of Things for Smart City, University of Macau. His research interests include modeling

and optimization of connected and autonomous vehicles and urban mobility in smart cities.



Haicheng Liao received the B.S. degree in software engineering from the University of Electronic Science and Technology of China in 2022 and the Ph.D. degree from the University of Macau. His research interests focus on autonomous driving.



Zhenning Li received his Ph.D. in Civil Engineering from the University of Hawaii at Manoa, Honolulu, Hawaii, USA, in 2019. He is currently an Assistant Professor at the State Key Laboratory of Internet of Things for Smart City, University of Macau, Macau. Over his academic career, he has published over 100 papers. His research focuses on the intersection of connected autonomous vehicles and Big Data applications in urban transportation systems. He has been honored with several awards, including the TRB best young researcher award and the CICTP best paper award, amongst others.