Scale-adaptive UAV Geo-localization via Height-aware Partition Learning

Quan Chen
Jiaxing University
Jiaxing, China
chenquan@alu.hdu.edu.cn

Bolun Zheng✉
Hangzhou Dianzi University
Hangzhou, China
blzheng@hdu.edu.cn

Tingyu Wang
Hangzhou Dianzi University
Hangzhou, China
tingyu.wang@hdu.edu.cn

Rongfeng Lu
Hangzhou Dianzi University
Hangzhou, China
rongfeng-lu@hdu.edu.cn

Yu Liu
Tsinghua University
Beijing, China
liuyu_thu@mail.tsinghua.edu.cn

Zhedong Zheng
University of Macau
Macau, China
zhedongzheng@um.edu.mo

Abstract

UAV Geo-Localization faces significant challenges due to the drastic appearance discrepancy between drone-captured images and satellite views. Existing methods typically assume a consistent scaling factor across views and rely on predefined partition alignment to extract viewpoint-invariant representations through part-level feature construction. However, this scaling assumption often fails in real-world scenarios, where variations in drone flight states lead to scale mismatches between cross-view images, resulting in severe performance degradation. To address this issue, we propose a scale-adaptive partition learning framework that leverages known drone flight height to predict scale factors and dynamically adjust feature extraction. Our key contribution is a height-aware adjustment strategy, which calculates the relative height ratio between drone and satellite views, dynamically adjusting partition sizes to explicitly align semantic information between partition pairs. This strategy is integrated into a Scale-adaptive Local Partition Network (SaLPN), building upon an existing square partition strategy to extract both fine-grained and global features. Additionally, we propose a saliency-guided refinement strategy to enhance part-level features, further improving retrieval accuracy. Extensive experiments validate that our height-aware, scale-adaptive approach achieves state-of-the-art geo-localization accuracy in various scale-inconsistent scenarios and exhibits strong robustness against scale variations. https://github.com/C-water/SaLPN

CCS Concepts

• **Computing methodologies** → **Visual content-based indexing and retrieval**.

Keywords

Geo-localization; Dynamic partition; Relative height ratio; Scale variations; Height-aware modulation; Cross-view matching



This work is licensed under a Creative Commons Attribution 4.0 International License.
MM '26, Rio de Janeiro, Brazil
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2213-4/2026/11
https://doi.org/10.1145/3767308.3835561

ACM Reference Format:
Quan Chen, Bolun Zheng, Tingyu Wang, Rongfeng Lu, Yu Liu, and Zhedong Zheng. 2026. Scale-adaptive UAV Geo-localization via Height-aware Partition Learning. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26), November 10–14, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3767308.3835561

1 Introduction

Unmanned Aerial Vehicles (UAVs), also known as drones, have gained popularity due to their ability to efficiently capture data with few occlusions and rich content. Differ from ground-views, UAVs can easily acquire multi-scale images by changing the flight altitude and camera parameters. Benefiting from these advantages, UAVs play an irreplaceable role in various fields, including agricultural operations [24], automatic driving [31] and aerial photography [7]. Regardless of the applications, precise positioning and navigation of drones are indispensable, which mainly rely on a global navigation satellite system (GNSS) and high-quality communication environment [79]. In scenarios with weak or failed signals, GNSS will incorrectly determine the geographic localization of drones. Therefore, vision-based UAV-view geo-localization (UVGL) develops a meaningful hot research, aiming to release the dependence on GNSS. As the key of this task, cross-view matching algorithms calculates the similarity between drone- and satellite-view images to achieve image matching. However, due to the appearance difference between drone-views and satellite-views, traditional image matching algorithms [1, 3, 41] are imprecise.

With the rapid development of deep learning, current methods [1, 10, 12, 18, 19, 30, 40, 49, 56, 57, 59, 62, 63, 69, 71, 72, 74] utilize pre-trained backbone networks to extract high-level semantic representations and perform image matching based on feature distances, achieving superior accuracy than traditional algorithms. To obtain robust discriminative representations, plentiful partition learning strategies have been proposed, which construct fine-grained features by segmenting high-level features to fully exploit contextual information. Depending on the pattern of feature segmentation, partition learning can be divided into two categories: **soft-partition** learning [13, 36, 37, 70, 77] and **hard-partition** learning [22, 23, 34, 58]. The former classifies features based on feature values to distinguish information categories, so the shape of partitions is uncontrollable. FSRA [13], as a representative of soft-partition learning, confirms that fine-grained features have a

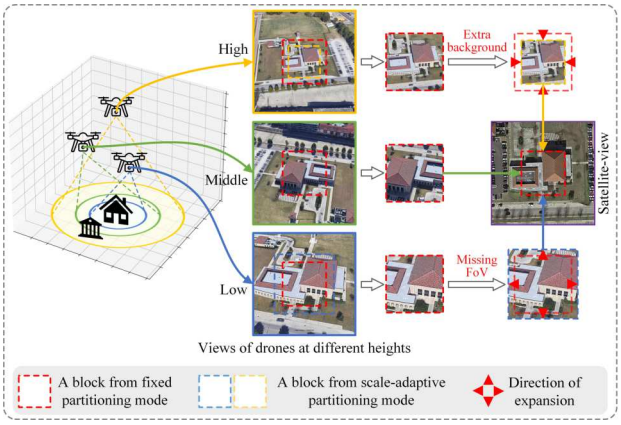

Figure 1: The simplified diagram of our research motivation. Given drone-view images captured from different heights (High-*yellow box*, Middle-*green box*, Low-*blue box*) and the corresponding satellite-view image (right).

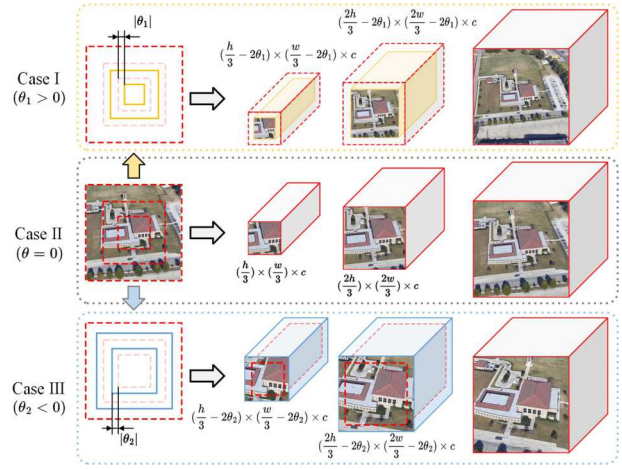

Figure 2: The proposed height-aware adjustment strategy ($N = 3$ for illustration).

consistent distribution with image content and can boost model performance. In contrast, inspired by the distribution similarity between drone- and satellite-views, hard-partition strategies set predefined templates with fixed shape to divide high-level features. As the pioneering hard-partition algorithm, LPN [58] designs non-overlapping square-ring templates to build fine-grained features. Subsequently, several studies [46, 78] further optimize the shape and number of templates to facilitate discriminative representation. Benefiting from the explicit utilization of contextual information that prevents the network from falling into a local optimum, hard-partition strategies reveal superior accuracy and compatibility than soft-partition strategies under same settings. Nevertheless, hard-partition strategies are more sensitive to spatial changes of input images. During the data collection process, the variations of drone flight state will lead to the scale mismatch of cross-view images. The predefined segmentation templates constrain misaligned semantic information to same parts, thus interfering with feature matching.

As shown in Fig. 1, we depict the limitation of hard-partition strategies in dealing with scale-misaligned image pairs. Taking LPN [58] as an example, the partition pairs contain similar semantic information in the case of similar cross-view scales (see drone-view with green box). When drone height changes, partition pairs of fixed shape contain misaligned semantic information, defined as imperfect matching (see drone-view with yellow or blue box). Following the perspective projection principle ("closer objects appear larger" in imaging), we argue that adjusting the partition size based on the shooting height of the drone can ensure the content consistency of partition pairs, thus boosting the matching of cross-view partition pairs. Concretely, for high-height drone-view, indent templates can reduce redundancy in the part-level representations, while expanded templates contributes to the content supplement of partitions for low-height drone-views. Imitating the process mentioned above, we propose a scale-adaptive partition learning framework named SaLPN, to enhance the content consistency of partition pairs when the shooting height of drone changes. Specifically, we design a square partition strategy (SPS) as shown in Case (II) of Fig. 2, which has two typical advantages: 1) multiple partitions constructed by SPS emphasize both fine-grained and global features, thereby exhibiting robustness against scale variations; 2) solid partitions have strong compatibility with subsequent algorithms. To promote the content consistency of partition pairs, we further propose a height-aware adjustment strategy (HAAS). We construct a scale factor θ based on the relative height ratios between drone and satellite views, and inject it into SPS to dynamically adjust the partition size of drone-view features. As shown in Case (I) and Case (III) of Fig. 2, HAAS controls variations of partitions, both in direction and degree. Furthermore, we design a saliency-guided refinement strategy (SGRS) to identify targets and environments based on feature activation degrees. SGRS maintains robustness to scale changes while improving model matching accuracy. Extensive experiments on University-1652 [74] and SUES-200 [76] show that SaLPN achieves state-of-the-art geo-localization accuracy in various scale-inconsistent scenarios and exhibits excellent robustness against scale variations. Our main contributions are:

- We introduce a new scale-adaptive partition learning framework that harness the pre-acquired drone height to explicitly addresses scale mismatches in cross-view scenarios.
- We further develop a saliency-guided refinement strategy to enhance the discriminative ability of part-level features by merging the coordinate attention.
- Extensive experiments show that, in comparison to both hard-partition and soft-partition strategies, SaLPN achieves splendid retrieval performance, especially in scenes with inconsistent cross-view scales.

2 Related Work

2.1 Cross-view Geo-localization

Cross-view Geo-localization aims to recognize the geographic location of an input image based on satellite database, which mainly emphasizes two distinct matching tasks, *i.e.*, the matching of ground and satellite views, as well as the matching of drone and satellite views. Traditional methods [1, 3, 41]rely on hand-crafted feature, yet their precision is seriously constrained by appearance differences of cross-view images.

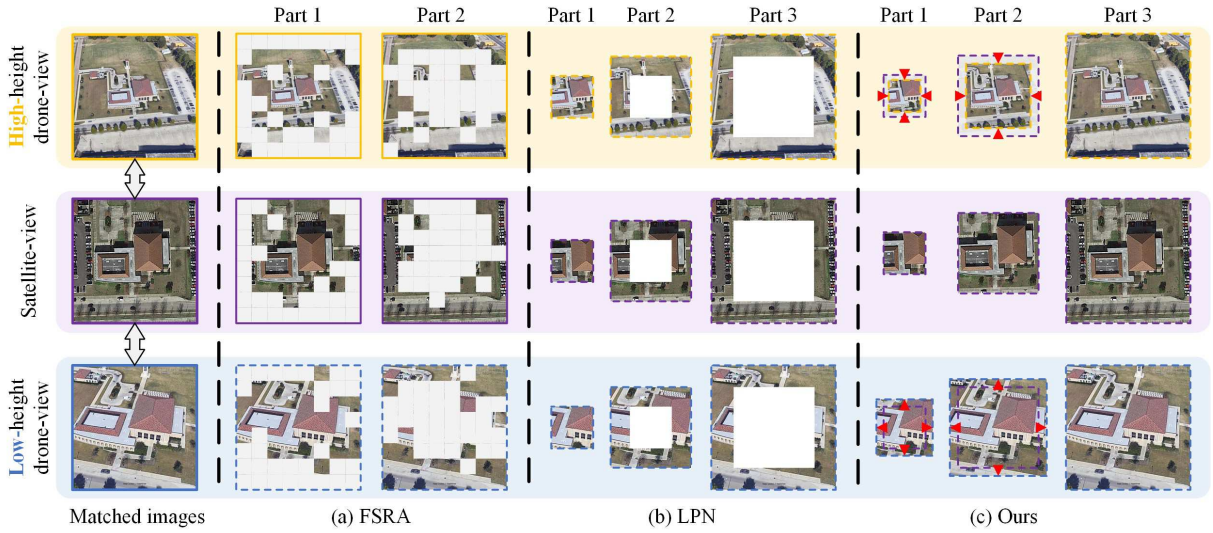

Figure 3: The comparison of our method with typical partition strategies.

With the growing popularity of UAV devices, several datasets [14, 29, 60, 76] introduce drone-view images to support for the UVGL task. For example, Zheng *et al.* [74] collect the University-1652 dataset, which contains pairs of ground-views, drone-views and satellite-views. Notably, Zheng *et al.* [74] introduce variations in both the shooting angle and shooting height of drone-views, thus enhancing the diversity of University-1652. Mainstream learning based solutions [27, 35, 38, 40, 42, 49, 51, 59–61, 64, 67, 74] converge upon a similar principle: mapping multi-view images into a shared high-dimensional space, subsequently facilitating image matching through similarity calculation. Several studies [6, 13, 34] prove that stronger backbone networks can significantly improve the accuracy of UVGL. Sun *et al.* [51] design a specific feature extractor named F3-Net to capture both local and global information. RK-Net [40] inserts unit subtraction attention modules inside the backbone to detect representative keypoints, yielding robust features against viewpoints. Considering the color discrepancy between drone-views and satellite-views, recent methods [53, 55] adjust the color distribution of drone-views to reduce domain gaps. Some efforts [27, 70] improve the optimization function in the training phase to facilitate domain alignment of cross-view images. To obtain robust semantic representations, plentiful partition learning strategies are also introduced, which will be detailed in next subsection. In this paper, we propose a UAV-view geo-localization algorithm applicable to scenes with inconsistent cross-view scales.

2.2 Part-based Representation Learning

Fine-grained features guide models to capture more comprehensive information, which has been proven effective in many fields, such as image recognition [8, 20, 47, 48] and segmentation [11, 25, 52, 68]. Inspired by classical part-based descriptors [2, 43, 45], numerous works in person re-identification [32, 50, 54, 65, 73, 75] develop local feature representation by introducing body structure information.

Previous research has laid the foundation for partition learning in UVGL, which can be divided into two categories based on the pattern of feature segmentation: **soft-partition** learning [13, 36, 37, 70] and **hard-partition** learning [9, 22, 23, 33, 34, 58]. Soft-partition strategies, represented by FSRA [13], leverage feature values to distinguish categories of semantic information. Subsequently, TransFG [70] adopts the second-order gradient of features as the partition index, yielding a decent performance boost. FSRA and TransFG divide features equally based on the setting that each part has equal area, resulting in misalignment of cross-view partition pairs in both large-scale and small-scale scenes. To address this issue, GeoFormer [36] utilizes a k-modes clustering algorithm to divide features adaptively. However, the above soft-partition strategies rely on the transformer framework [17]. Drawing upon the similarity between drone- and satellite-views, some methods employ fixed templates to generate fine-grained features, which are defined as hard-partition strategies. LPN [58] designs non-overlapping square-ring templates to segment features, thus mining contextual information. Based on LPN, several studies [22, 46, 78] optimize the shape and number of templates to further facilitate discriminative representation. Aiming at the problem of position offset, SDPL [6] proposes a shifting-fusion strategy, which adjusts the segmentation center to align partitions with non-centered targets. The hard partition strategies exhibit commendable accuracy and compatibility, while being sensitive to cross-view scale variation. Recently, Safe-Net [39] proposes a saliency-guided self-adaptive partition strategy, which achieves preferable results than previous methods in scale inconsistency scenarios. However, Safe-Net [39] also produces significant performance degradation in extreme scenarios. We intend to design an robust algorithm against scale variations.

To visually reveal the dependence of partition learning on cross-view scales, we illustrate two representative methods (FSRA [13] and LPN [58]) for dealing with high-altitude scenes and low-altitude scenes, as shown in Fig. 3. Limited by the setting of equal partition area, FSRA divides features of different categories into the same partition, resulting in feature misalignment. LPN employs fixed templates to generate fine-grained features. Notably, in scenarios where cross-view scales are inconsistent, partition pairs contain misaligned semantic content. In this paper, we study a part-based representation learning for UVGL. In comparison with

soft-partition strategies, our method achieves superior performance and compatibility, whereas, in contrast to hard-partition strategies, it exhibits enhanced robustness against scale variations.

3 Methodology

Task definition. The geo-localization dataset contains image pairs captured from different platforms, defined as $\{x_D, x_S\}$, where subscripts D and S indicate drone and satellite-platform, respectively. We denote the label as $y \in [1, C]$, where C indicates the number of categories. For instance, the label of the University-1652 [74] dataset is $y \in [1, 701]$. For UAV-view geo-localization, we intend to learn a mapping function that projects images from various platforms to one shared semantic space. Image pairs with same labels are close to each other, and vice versa are far away. Therefore, given the drone-view images, user could quickly search the corresponding images from other platform during the inference time. According to the meta-data of other platform, we localize the drone even in the GNSS-denied areas.

Overview. As depicted in Fig. 4, SaLPN consists of three stages: feature extraction, scale-adaptive partition learning and classification supervision. In the feature extraction stage, a backbone network maps drone- and satellite-view images to high-level feature maps. Then, we introduce scale-adaptive partition learning to generate fine-grained features with consistent content. We further propose a saliency-guided refinement strategy to classify each part-level feature, thus obtaining finer discriminative representations. Finally, we adopt a classifier to predict the category of features obtained from different views, followed by using cross-entropy loss to optimize the model.

3.1 Feature Extraction

Both drone-view and satellite-view images fall within the category of aerial-view images, exhibiting comparable feature domains. Consequently, we employ one feature extractor to process cross-view images. As mentioned before, SaLPN belongs to hard-partition learning and is compatible with various backbone networks, such as ResNet [26] and ViT [17]. Following previous works [6, 58], we adopt ResNet-50 for illustration. Given an image $x_i \in \mathbb{R}^{3\times512\times512}$, after the backbone network and an upsampling layer, we can acquire the corresponding high-level feature map $f_i \in \mathbb{R}^{2048\times128\times128} (i \in \{D, S\})$. The subscripts D and S refer to the drone- and satellite-view. We emphasize that the degree of variation in partition size can be refined by upsampling feature maps, so we introduce a $\times4$ up-sampling layer $\mathcal{U}_{\times4}$.

3.2 Scale-adaptive Partition Learning

To capture aligned contextual information even in scale-inconsistent scenes, we propose a scale-adaptive partition learning, consisting of a square partition strategy, a height-aware adjustment strategy, and a saliency-guided refinement strategy.

Square partition strategy (SPS). We propose a simple yet effective partition strategy to generate fine-grained features. As shown in Case(II) of Fig. 2, SPS splits the input feature into N parts following square templates. Concretely, after SPS processing, the high-level feature map $f_i \in \mathbb{R}^{2048\times128\times128}$ is segmented into N part-level feature $f_i^n \in \mathbb{R}^{2048\times\frac{128n}{N}\times\frac{128n}{N}} (n \in [1, N])$, where n denotes the n-th part. The conventional SPS processing can be formulated as:

$$\Omega(f_i^n) = \{f_i^n | \mathcal{F}_{\mathcal{SPS}}(f_i; N)\}, \tag{1}$$

where $\mathcal{F}_{\mathcal{SPS}}(\cdot)$ stands for the process of SPS, and $\Omega(f_i^n)$ denotes the set of all part-level features.

Height-aware adjustment strategy (HAAS). Existing hard-partition methods adopt uniform templates for processing multi-view images. However, in scenarios where drone-view scale undergo significant variations, these methods exhibit notable performance degradation. The primary reason is the misalignment of semantic content within partition pairs. To alleviate this limitation, we propose an innovative height-aware adjustment strategy, *i.e.*, dynamically adjusting partition size in response to drone-view scale changes, thus boosting content consistency between partition pairs.

First, we assume that the shooting height of drone-views is acquirable, while the height of satellite images can be estimated by comparing the area per unit pixel with a fixed height drone-view image. We declare that this content scale ratio estimation is in accordance with imaging principles. To quantify the change in partition size, we further construct a scale factor θ based on the relative height between cross-view images:

$$\theta = \text{round}(\frac{H_D - H_S}{H_S} \cdot \alpha), \tag{2}$$

where H_D and H_S denote the height from drone-view and satellite-view to the ground, respectively; round$(\cdot)$ stands for the rounding operation, and partition adjustment factor α denotes a hyper-parameter utilized to regulate the degree of partition change. (θ>0) indicates that the partition needs to be decreased, and (θ<0) vice versa. Note that the height of satellite-view H_S is constant, which is the regular setting for UVGL.

Then, different from the traditional SPS in Eq. 1, we further take θ into consideration, which adjusts partitions of drone-views:

$$\Omega(f_D^n) = \{f_D^n | \mathcal{F}^{\mathcal{D}}_{\mathcal{HAAS}}(f_D; N; \theta)\}, \tag{3}$$

where $\mathcal{F}^{\mathcal{D}}_{\mathcal{HAAS}}(\cdot)$ denotes the process of HAAS specific for drone-views. For ease of understanding, we depict the $\mathcal{F}^{\mathcal{D}}_{\mathcal{HAAS}}$ process in two cases (*i.e.*, θ>0 and θ<0), as shown in Fig. 2 (Case I and III). We employ θ to adjust the size of partitions, and the shape of each partition pair is expressed as follows:

$$\begin{cases} f_D^n \in \mathbb{R}^{2048\times(\frac{128n}{N}-2\theta)\times(\frac{128n}{N}-2\theta)}, \\ f_S^n \in \mathbb{R}^{2048\times(\frac{128n}{N})\times(\frac{128n}{N})}. \end{cases} \tag{4}$$

In practice, there are two extreme cases that should be emphasized: 1) when θ>0, the size of the smallest partition must be positive; 2) while θ<0, the minimum partition block expands into global features, *i.e*, all partitions are equal to the global feature. This means that α in Eq. 2 has a limited range when the height range of the dataset is known. Next, we will derive the range of values of α.

Specifically, when θ>0 (*i.e*, H_D>H_S), we have $1 \le \frac{128}{2N} - \theta$ to avoid non-positive partitions. Based on Eq. 2, we can derive the range of the hyperparameter α:

$$0 \le \alpha \le \frac{64 - N}{N} \cdot \frac{H_S}{H_D^{max} - H_S}, \tag{5}$$

where H_D^{max} denotes the maximum height of drone-views.

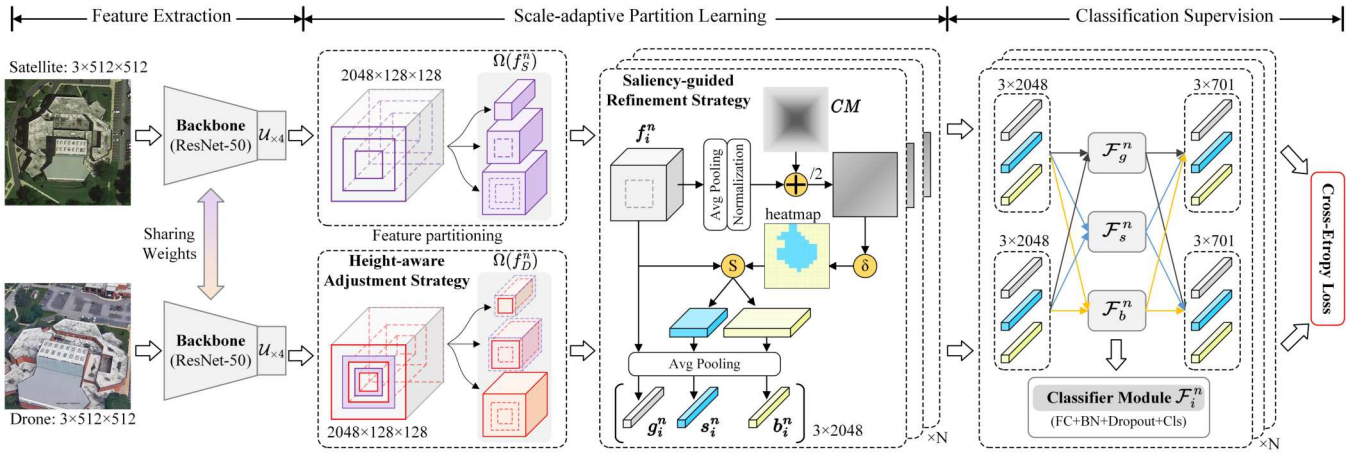

Figure 4: Overview of the SaLPN framework. Cross-view features are dynamically partitioned for scale alignment based on the drone's flight height, and then enhanced by a saliency-guided refinement strategy to generate robust part-level descriptors. ⊕ denotes the element-wise addition; ⓑ denotes threshold-based binarization; Ⓢ denotes the feature separation.

Conversely, when $\theta<0$ (*i.e*, $H_D<H_S$), we have $\frac{128}{2N}-\theta \leq \frac{128}{2}$. Based on Eq. 2, we also can derive the range of the hyperparameter α:

$$0 \leq \alpha \leq \frac{64(N-1)}{N} \cdot \frac{H_S}{H_S - H_D^{min}}, \tag{6}$$

where H_D^{min} denotes the minimum height of drone-views.

If all same-height image pairs have $H_S = H_D$, we could derive $\theta = 0$, which is independent from the selection of α. Therefore, for a dataset with a clear height range, the value of α needs to satisfy both Eq. 5 and 6.

Saliency-guided refinement strategy (SGRS). Each solid partition inherently contains both target and environment information. To further refine part-level features, we design a saliency-guided refinement strategy. The overall process is depicted in Fig. 4, and can be expressed as:

$$(g_i^n, s_i^n, b_i^n) = \mathcal{F}_{\mathcal{SGRS}}{}^n(f_i^n), \tag{7}$$

where $\mathcal{F}_{\mathcal{SGRS}}{}^n(\cdot)$ denotes the process of SGRS, and g_i^n, s_i^n, b_i^n are denote the output global, salient and background feature, respectively.

Specifically, assuming the input partition feature is $f_i^n \in \mathbb{R}^{2048\times h\times w}$, we first generate the heatmap $heat_i^n \in \mathbb{R}^{1\times h\times w}$ to reflect importance regions of the input. Considering that targets are usually centered, we introduce an auxiliary coordinate map (CM) to direct the heatmap's focus towards the central region. CM indicates inverse pixel-to-center distance, ranging from 1 (center) to 0 (farthest). To be consistent with the partition shape, we adopt Chebyshev distance to generate CM:

$$CM(u,v) = max(\frac{h}{2}, \frac{w}{2}) - max(|u - \frac{h}{2}|, |v - \frac{w}{2}|), \tag{8}$$

where (u, v) denotes the coordinates of pixels. The generation of $heat_i^n$ can be formulated as:

$$heat_i^n(u,v) = (Norm(Avg(f_i^n(u,v))) + CM(u,v))/2, \tag{9}$$

where $Norm(\cdot)$ denotes the normalization operation, and $Avg(\cdot)$ denotes the average pooling operation. Next, we set a threshold δ to split the heatmap, resulting in a binary mask $Mask_i^n$:

$$Mask_i^n(u,v) = \begin{cases} 1, & heat_i^n(u,v) \geq \delta, \\ 0, & heat_i^n(u,v) < \delta. \end{cases} \tag{10}$$

Finally, in reference to FSRA [13], we leverage the value of $Mask_i^n$ to split the input feature f_i^n, which can be expressed as follows:

$$\begin{cases} s_i^n = Avg(f_i^n(u,v) \odot Mask_i^n(u,v)), \\ b_i^n = Avg(f_i^n(u,v) \odot (1 - Mask_i^n(u,v))), \\ g_i^n = Avg(f_i^n), \end{cases} \tag{11}$$

where $\odot$ denotes the pixel-level multiplication, $s_i^n \in \mathbb{R}^{2048\times1\times1}$ and $b_i^n \in \mathbb{R}^{2048\times1\times1}$ denote the salient feature and background feature, respectively. We also extract $g_i^n \in \mathbb{R}^{2048\times1\times1}$ to preserve the input information. It is worth mentioning that SGRS is implemented for each part-level features of both drone-view and satellite-view, resulting in $3N$ feature vectors.

3.3 Classification Supervision

We employ $3N$ classifier modules to map feature vectors of all sources into three corresponding shared space. As shown in Fig. 4, the classifier module includes a fully connected layer (FC), a batch normalization layer (BN), a dropout layer (Dropout), and a classification layer (Cls). The process can be expressed as:

$$p_g^{i,n} = \mathcal{F}_g^n(g_i^n), \quad p_s^{i,n} = \mathcal{F}_s^n(s_i^n), \quad p_b^{i,n} = \mathcal{F}_b^n(b_i^n), \tag{12}$$

where $\mathcal{F}_g^n, \mathcal{F}_s^n, \mathcal{F}_b^n$ are three independent classifier modules, respectively. It is worth noting that the classifiers adopted for different granularity feature vectors are also independent. Subsequently, we adopt a cross-entropy loss function to optimize our framework, with the objective of minimizing the distance between features of the same geo-tag:

$$\mathcal{L}_{CE} = \sum_{i,n} -log(p_g^{i,n}(y)) - log(p_s^{i,n}(y)) - log(p_b^{i,n}(y)), \tag{13}$$

where $p_g^{i,n}(y)$ denotes the logit score of the ground-truth geo-tag y based on the global feature g_i^n. Similarly, $p_s^{i,n}(y)$ and $p_b^{i,n}(y)$ are the predicted probability of the ground-truth location based on the salient and background feature, respectively.

4 Experiment

4.1 Datasets

The main experiments are performed based on the University-1652 [74] dataset for the following reasons: 1) University-1652 contains diverse scenarios, including 1652 university buildings worldwide, where each building comprises 1 satellite-view image and 54 drone-view images. The training set includes 701 buildings of 33 universities, and the other 951 buildings of the remaining 39 universities are classified as the test set. 2) drone-view images exhibits considerable scale variation, with height ranging from 123.5m to 256m (*i.e.*, $H_D \in [123.5, 256]$), supporting the verification of the anti-scale robustness of SaLPN. The test set is equally divided into three parts according to the shooting height of drone-view: **Long**, **Middle**, and **Short**. Then, comparing satellite-view images with drone-view images, we set the value of H_S to 189.75m, which corresponds to the midpoint of the drone height range.

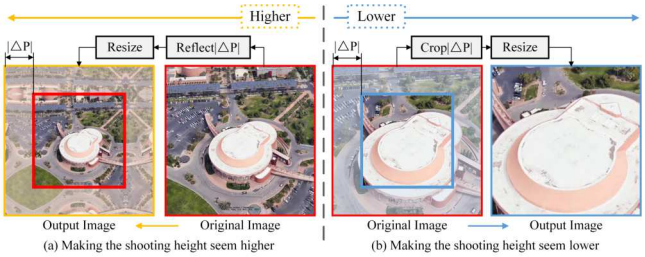

Figure 5: The data augmentation strategy to extend the shooting height range of drone-views from the visual perspective. (a) Given a drone image, mirror a strip ΔP pixels wide around the input image and resize it to initial resolution; (b) Given a drone image, crop out a square ring ΔP pixels wide around the input image and resize it to initial resolution.

To further expand the height range of drone-views, we design a data augmentation strategy to simulate more extreme settings. As shown in Fig. 5, we introduce a parameter ΔP to adjust image scales (equivalent to drone height), which can be expressed as: (ΔP>0) means that making the height of drone-view seem higher (see Fig. 5(a)), while (ΔP<0) means that making the height of drone-view seem lower (see Fig. 5(b)). $|\Delta P|$ denotes the degree of scale adjustment. Then, we roughly estimate the corresponding new height of synthesized images with ΔP as:

$$H'_D = H_D + \lambda_{aug} \cdot \Delta P, \tag{14}$$

where λ_{aug}=0.7 is the empirical value, estimated by the visual scale and height of the existing drone-view images. By applying the aforementioned strategy to **Long** and **Short** sets, we generate test sets with extreme scales, named as **Long** (ΔP>0) and **Short** (ΔP<0), respectively. Note that drone-views within **Long** (ΔP>0) exhibit higher heights (meaning smaller image scale), whereas those within **Short** (ΔP<0) have lower heights (meaning larger image scale).

The SUES-200 [76] dataset collects drone-view images for 200 scenes at four heights (*i.e.*, 150m, 200m, 250m, 300m) independently, where each scene contains 50 drone-view images and 1 satellite-view image. The training set includes 120 buildings, and the other 80 buildings are classified as the test set. The training and testing of models are performed with data of the same height.

4.2 Implementation Details

For the SaLPN framework, we set the number of partitions to N=4, and partition adjustment factor α=14.0 by default. For SRGS, we adopt Chebyshev distance to generate coordinate map CM, and set the heatmap segmentation threshold δ=0.5. For the University-1652 dataset, we roughly set drone-views height $H_D \in [123.5, 256]$ and satellite-views height H_S=189.75m, respectively. For the SUES-200 dataset, we roughly set satellite-views height H_S=300m and the drone view altitude to 150/200/250/300m, respectively. Training details are included in the supplementary materials

4.3 Comparison with the State-of-the-art

Performance on University-1652. Except for the standard test set provided, we employ Eq. (14) to generate two supplementary test sets, named **Long** (ΔP=100) and **Short** (ΔP=-100), which exhibit more extreme height of drone-views (*i.e.*, more significant cross-view scale differences). As shown in Table 1, the quantitative results are divided into two groups, ResNet-based and ViT-based methods. On the ResNet-based track, SaLPN achieves 84.37% R@1 and 86.51% AP on default test set, 64.89% R@1 and 68.86% AP on **Short** (ΔP=-100) set and 64.08% R@1, and 68.75% AP on **Long** (ΔP=100) set, respectively. On the default test set, SaLPN achieves competitive performance with recent SDPL [6] and clearly outperforms other solutions, such as D^2-GeM [56] and Sample4Geo [15]. Moreover, it is remarkable that SaLPN yields superior results on both the **Short** (ΔP=-100) and **Long** (ΔP=100) compared with partition-based LPN [13] and SDPL [6]. Specifically, compared with LPN, SaLPN improves R@1 from 47.24%, 35.57% to 64.89% (+17.65%), 64.08% (+28.51%), respectively. On the ViT-based track, SaLPN is compatible with the ViT framework and has surpassed the sate-of-the-art methods on **Short** (ΔP=-100) and **Long** (ΔP=100), achieving an average R@1 and AP of 79.69% and 82.46%. This shows that SaLPN is robust to scenes with inconsistent cross-view scales.

Performance on SUES-200. To further validate the effectiveness of SaLPN, we report the results on SUES-200 [76]. As shown in Table 2, SaLPN achieves competitive results under different height. Compared with FSRA [13] at four heights, SaLPN boosts the R@1 from 68.25%, 83.00%, 90.68%, 91.95% to 81.92% (+13.67%), 87.90% (+4.90%), 92.78% (+2.10%), 93.03% (+1.08%), respectively.

Qualitative Results. we visualize some retrieved results in Fig. 6. To prove the anti-scaling robustness of SaLPN, we provide same numbered results generated by baseline [26] and LPN [58] on different test sets. For the original test set, both LPN and SaLPN are able to retrieve the correct satellite labels. when the drone-view images undergo more drastic scale variations, it can be observed that SaLPN still achieves correct results. This benefits from the fact that SaLPN explicitly aligns contents of partition pairs, thus mining robust contextual information.

4.4 Ablation Studies and Further Discussion

Robustness of SaLPN against scale-inconsistent scenarios. As a hard-partition-based method, our method can alleviate the dependence on scale consistency of cross-view images by dynamically adjusting the size of partition pairs. To fully verify the robustness of SaLPN against scale-inconsistent scenarios, we first synthesize a series of test sets with lower and higher height of drone-views,

Table 1: Comparison results on University-1652. Since the code for method marked with ** is not publicly available, its performance in extreme scenarios could not be evaluated.

Methods	Publication	Backbone	Standard Test set		Short (ΔP=-100)		Long (ΔP=100)		Average	
			R@1	AP	R@1	AP	R@1	AP	R@1	AP
Instance loss [74]	ACMMM'20	ResNet-50	60.16	64.98	37.62	43.36	23.12	28.88	40.30	45.74
LPN [58]	TCSVT'21	ResNet-50	81.48	84.06	47.24	52.93	35.57	41.51	54.76	59.50
CA-HRS [44]	ACCV'22	ResNet-50	81.00	83.64	48.55	54.21	34.94	40.94	54.83	59.59
MCCG [49]	TCSVT'23	ResNet-50	70.94	75.04	34.54	40.60	25.06	31.25	43.51	48.96
Sample4Geo [15]	ICCV'23	ResNet-50	78.62	82.11	53.30	59.05	41.03	47.73	57.65	62.96
D^2-GeM [56]	SPL'24	ResNet-50	84.49	86.81	55.55	60.92	37.30	43.48	59.11	63.73
DWDR(w.BT) [59]	TGRS'24	ResNet-50	71.79	75.71	49.16	54.75	37.24	43.90	52.73	58.12
DWDR(w.LPN) [59]	TGRS'24	ResNet-50	83.18	85.63	51.29	56.71	36.36	42.71	56.94	61.68
SDPL [6]	TCSVT'24	ResNet-50	85.19	87.43	60.24	65.10	48.29	54.32	64.57	68.95
**Chen *et al.* [4]	JSTAR'25	ResNet-50	77.67	80.79	–	–	–	–	–	–
**SCPNet [21]	TCSVT'25	ResNet-50	79.96	83.04	–	–	–	–	–	–
**ECSNet [66]	TGRS'26	ResNet-50	70.92	74.78	–	–	–	–	–	–
SaLPN	–	ResNet-50	84.37	86.51	64.89	68.86	64.08	68.75	71.11	74.71
Instance loss [74]	ACMMM'20	ViT-S	80.34	83.05	68.01	71.83	61.23	65.99	69.86	73.62
LPN [58]	TCSVT'21	ViT-S	87.07	88.91	69.62	73.48	69.81	73.98	75.50	78.79
FSRA [13]	TCSVT'21	ViT-S	85.10	87.19	71.48	75.17	65.67	70.04	74.08	77.46
SDPL [6]	TCSVT'24	ViT-S	85.57	87.61	68.90	72.76	70.56	74.65	75.01	78.34
**MMHCA [5]	CJA'25	ViT-S	76.25	79.39	–	–	–	–	–	–
**QDFL [28]	TGRS'25	ViT-S	86.46	88.54	–	–	–	–	–	–
Safe-Net [39]	TIP'25	ViT-S	88.30	89.93	73.32	76.80	73.86	77.51	78.49	81.41
SaLPN	–	ViT-S	88.19	89.79	73.45	76.91	77.44	80.69	79.69	82.46

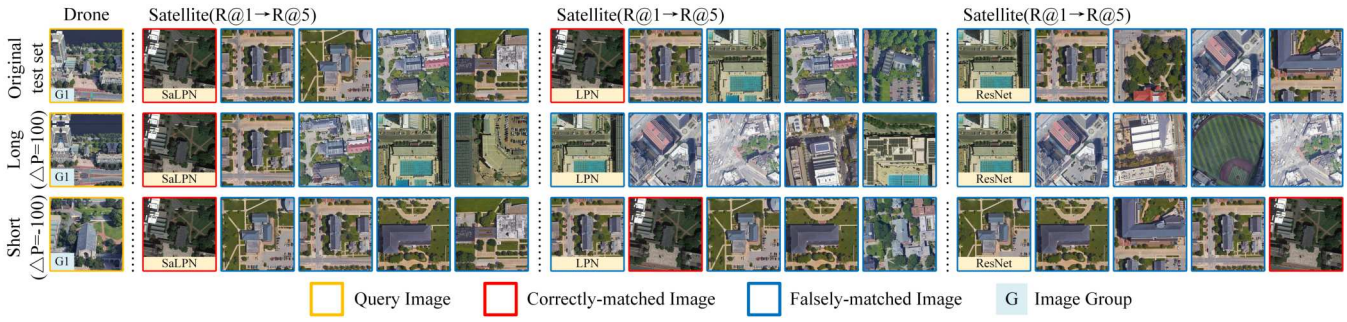

Figure 6: Image retrieval results obtained with SaLPN, LPN and ResNet. Specifically, we present the correspondingly numbered image retrieval outcomes from the original test set (a), Long (ΔP=100) (b), and Short (ΔP=-100) (c), respectively.

Table 2: Comparison results on SUES-200.

Methods	150m		200m		250m		300m	
	R@1	AP	R@1	AP	R@1	AP	R@1	AP
SUES-200 [76]	55.65	61.92	66.78	71.55	72.00	76.43	74.05	78.26
LCM [16]	43.42	49.65	49.42	55.91	54.47	60.31	60.43	65.78
LPN [58]	61.58	67.23	70.85	75.96	80.38	83.80	81.47	84.53
FSRA [13]	68.25	73.45	83.00	85.99	90.68	92.27	91.95	93.46
SaLPN	81.92	85.23	87.90	90.02	92.78	94.24	93.03	94.37

named **Short** (ΔP<0) and **Long** (ΔP>0), where a larger $|\Delta P|$ represents a larger drone-view scale variation. The results are shown in Table 3. On the ResNet50-based track, SaLPN shows the best retrieval accuracy, with significantly less performance degradation than other partition-based methods. In particular, compared with LPN and SDPL, SaLPN improves average R@1 from 56.53%, 65.02% to 70.66% (+14.13%, +5.64%), respectively. Furthermore, the integration of HAAS into SaLPN led to a notable improvement in retrieval accuracy in extreme scenarios, with a 5.42% improvement in the average R@1. This highlights that HAAS facilitates alignment of fine-grained features from multiple views. On the ViT-based track, SaLPN also achieves optimal retrieval accuracy across datasets of all scales, with an average R@1 of 78.97%. Compared with FSRA and LPN, SaLPN improves average R@1 by 5.11% and 3.88%. This fully validates the robustness of SaLPN to cross-view scale variations, as well as the compatibility with various backbone networks.

Effectiveness of primary components. The core components of SaLPN are SPS, HAAS and SGRS. We construct five derived models to validate the effectiveness of each strategy, and results are shown in Table 5. Firstly, fine-grained partitions generated by SPS contain both local and global features, thus learning rich contextual information. Compared to the baseline, SPS provides +19.76% R@1, +21.60% R@1 and +24.88% R@1 on three test sets. When HAAS is combined with SPS, the model's accuracy is notably enhanced in scenarios involving scale inconsistency, while maintaining stable performance in the standard test set. For instance, after removing HAAS from SaLPN, the R@1 of model (SPS+SGRS) decreases by

Table 3: Ablation study on robustness of SaLPN to scale-inconsistent scenarios. we reported the R@1 accuracy for baseline, LPN, FSRA, SDPL and SaLPN on various test sets. The experimental results are divided into two groups: ResNet50-based at the top and ViT-based at the bottom. best and second best performance are in red and blue colors, respectively.

Test sets	**Short** ($\Delta P \leq 0$)						**Middle**	**Long** ($\Delta P \geq 0$)						Average
	-150	-100	-75	-50	-25	0	–	0	+25	+50	+75	+100	+150	
Instance loss [74]	9.26	37.62	48.40	56.19	61.27	64.81	63.58	52.07	40.40	33.82	27.88	23.12	16.25	41.12
FSRA [13]	7.39	33.24	46.18	54.51	59.87	64.15	65.07	48.41	36.50	29.43	24.53	20.95	26.51	39.74
LPN [58]	9.73	47.24	64.19	74.82	80.49	84.11	86.30	74.01	59.38	49.58	41.20	35.57	28.31	56.53
SDPL [6]	19.36	60.24	72.30	79.25	83.85	86.39	88.56	80.58	69.92	62.25	54.76	48.29	39.57	65.02
SaLPN (w/o HAAS)	20.56	60.08	72.33	78.89	82.52	84.75	86.27	79.63	70.77	64.24	57.63	51.13	39.34	65.24
SaLPN	25.68	64.89	74.95	80.20	83.54	85.09	86.16	81.88	75.95	72.24	68.47	64.08	55.48	70.66
Instance loss [74]	41.06	68.01	74.39	77.96	79.77	80.49	81.87	78.66	72.98	69.16	65.73	61.23	54.45	69.67
FSRA [13]	39.35	71.48	78.71	82.94	84.44	85.45	86.39	83.44	78.86	74.73	70.72	65.67	57.26	73.80
LPN [58]	32.93	69.62	78.32	83.13	85.97	87.11	88.41	85.71	81.33	78.12	73.80	69.81	62.01	75.09
SDPL [6]	33.93	68.90	77.22	82.07	84.62	85.45	86.99	84.27	80.27	77.59	74.24	70.56	63.72	74.60
Safe-Net [39]	38.96	73.32	80.85	85.51	87.69	87.91	89.36	87.64	84.70	81.68	78.09	73.86	66.89	78.18
SaLPN (w/o HAAS)	38.79	73.66	80.66	84.39	86.59	87.34	89.01	87.46	84.05	81.13	78.01	74.66	66.20	77.84
SaLPN	37.98	73.45	80.82	85.01	87.13	87.69	89.16	87.73	84.97	82.83	80.48	77.44	71.96	78.97

Table 4: Ablation study on different variants of SGRS.

Output features			Coordinate maps			Standard Test set		**Short** (ΔP=-100)		**Long** (ΔP=100)		Average	
g_i	s_i	b_i	Euclidean	Manhattan	Chebyshev	R@1	AP	R@1	AP	R@1	AP	R@1	AP
✓						80.69	83.35	59.75	64.17	60.71	65.56	67.05	71.02
✓	✓				✓	82.15	84.57	64.36	68.36	62.00	66.72	69.50	73.21
✓		✓			✓	83.56	85.88	62.37	66.70	61.55	66.36	69.16	72.98
	✓	✓			✓	83.99	86.14	63.57	67.82	61.49	66.19	69.68	73.38
✓	✓	✓				79.66	82.42	56.38	61.51	57.87	62.72	64.63	68.88
✓	✓	✓	✓			84.02	86.20	64.16	68.40	63.16	67.69	70.44	74.09
✓	✓	✓		✓		82.69	85.05	59.15	63.73	60.76	65.71	67.53	71.49
✓	✓	✓			✓	84.37	86.51	64.89	68.86	64.08	68.75	71.11	74.71

Table 5: performance comparison of SaLPN framework with different components. ✓ indicates that the module is applied.

Model structure			Standard Test set		**Short** (ΔP=-100)		**Long** (ΔP=100)	
SPS	HAAS	SGRS	R@1	AP	R@1	AP	R@1	AP
			60.16	64.98	37.62	43.36	23.12	28.88
✓			79.92	82.58	59.22	63.80	48.00	53.64
✓	✓		80.69	83.35	59.75	64.17	60.71	65.56
✓		✓	83.54	85.84	60.08	64.62	51.13	56.98
✓	✓	✓	84.37	86.51	64.89	68.86	64.08	68.75

0.83%, 4.81%, and 12.95% on standard test set, **Short** (ΔP=-100) and **Long** (ΔP=100), respectively. We emphasize that drone-views with low height contain incomplete depictions of targets and limited environmental information. consequently, it is plausible that the performance gain of HAAS is lower in the **Short** set compared to the **Long** set. To further improve the retrieval accuracy of SaLPN, we introduce SGRS to refine salient regions of partitions. It can be observed that SGRS yields 3.68%, 5.14% and 3.37% R@1 improvement over the model (SPS+HAAS).

Variants of SGRS. SGRS segments part-level features based on heatmaps to obtain global features, salient features and background features, which are matched independently. To fully verify the effectiveness of SGRS, we construct a variety of SGRS variants, including different output combinations as well as coordinate map *CM* (Diagram in supplementary materials). Table 4 shows the experimental results. Regardless of the specific combination of output modes employed by SGRS to refine partitions, it consistently outperforms the model without SGRS (*i.e.*, only g_i). On the other hand, compared to Euclidean distance and Manhattan distance, the utilization of Chebyshev distance within SGRS yields an optimal gain. This phenomenon can be attributed to the fact that *CM* generated by the Chebyshev paradigm exhibits a structural similarity to the square partition strategy employed. Since the feature segmentation process is unsupervised, the injected *CM* can assist SGRS in judging the importance of features and classifying neighborhood features into the same class. After removing of *CM* from SGRS, the average R@1 and AP metrics of SaLPN decreased by 6.48% and 5.83%.

5 Conclusion

In this paper, we investigate UAV-view geo-localization under scale-inconsistent scenarios and propose an effective part-based framework, termed SaLPN. Specifically, a square partition strategy is introduced to capture fine-grained local features while preserving the global structure. To mitigate scale inconsistency across views, we design a height-aware adjustment strategy that adapts the partition size of the drone branch based on the height ratio, improving the consistency of corresponding partitions. In addition, a saliency-guided refinement strategy is employed to enhance discriminative feature learning while preserving robustness to scale variations. Extensive experiments and ablation studies on University-1652 and SUES-200 demonstrate the effectiveness, robustness, and backbone compatibility of the proposed method. In future work, we will construct a real-world dataset with drone pose information to support UAV geo-localization and self-positioning.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62371175; 62501221).

References

[1] Mayank Bansal, Harpreet S Sawhney, Hui Cheng, and Kostas Daniilidis. 2011. Geo-localization of street views with aerial image databases. In *Proceedings of the 19th ACM international conference on Multimedia*. 1125–1128.

[2] Guillaume Bouchard and Bill Triggs. 2005. Hierarchical part-based visual object categorization. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 1. IEEE, 710–715.

[3] Francesco Castaldo, Amir Zamir, Roland Angst, Francesco Palmieri, and Silvio Savarese. 2015. Semantic cross-view matching. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 9–17.

[4] Junlong Chen, Gongjian Wen, Haojun Jian, and Xuxiang Fan. 2025. A Visual Localization Benchmark for UAVs in Complex Multi-Terrain Environments. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (2025), 1–15.

[5] Nanhua Chen, Taishan Lou, and Liangyu Zhao. 2025. MMHCA: Multi-feature representations based on multi-scale hierarchical contextual aggregation for UAV-view geo-localization. *Chinese Journal of Aeronautics* 38, 6 (2025), 103242.

[6] Quan Chen, Tingyu Wang, Zihao Yang, Haoran Li, Rongfeng Lu, Yaoqi Sun, Bolun Zheng, and Chenggang Yan. 2024. SDPL: Shifting-Dense Partition Learning for UAV-view Geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology* (2024), 1–1. doi:10.1109/TCSVT.2024.3424196

[7] Quan Chen, Zihao Yang, Bolun Zheng, Zhiming Cheng, Rongfeng Lu, Tingyu Wang, and Yu Liu. 2026. Learning Scene Priors from Satellite Imagery for UAV Image Motion Deblurring. *IEEE Transactions on Geoscience and Remote Sensing* (2026).

[8] Yue Chen, Yalong Bai, Wei Zhang, and Tao Mei. 2019. Destruction and construction learning for fine-grained image recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5157–5166.

[9] Yireng Chen, Zihao Yang, and Quan Chen. 2023. A cross-view matching method based on dense partition strategy for uav geolocalization. In *Proceedings of the 2023 workshop on uavs in multimedia: capturing the world from a new perspective*. 19–23.

[10] Zhongwei Chen, Zhao-Xu Yang, and Hai-Jun Rong. 2024. Multi-Level Embedding and Alignment Network with Consistency and Invariance Learning for Cross-View Geo-Localization. *arXiv preprint arXiv:2412.14819* (2024).

[11] Senlin Cheng, Haopeng Sun, Tao Xie, Hangyue Zhao, Yiqiang Chen, Bolei Xu, and XiaoBo Li. 2025. Spt: Sequence prompt transformer for interactive image segmentation. 5 pages.

[12] Meng Chu, Zhedong Zheng, Wei Ji, Tingyu Wang, and Tat-Seng Chua. 2024. Towards natural language-guided drones: Geotext-1652 benchmark with spatial relation matching. In *European Conference on Computer Vision*. Springer, 213–231.

[13] Ming Dai, Jianhong Hu, Jiedong Zhuang, and Enhui Zheng. 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 (2021), 4376–4389.

[14] Ming Dai, Enhui Zheng, Zhenhua Feng, Lei Qi, Jiedong Zhuang, and Wankou Yang. 2023. Vision-based UAV self-positioning in low-altitude urban environments. *IEEE Transactions on Image Processing* (2023).

[15] Fabian Deuser, Konrad Habel, and Norbert Oswald. 2023. Sample4geo: Hard negative sampling for cross-view geo-localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 16847–16856.

[16] Lirong Ding, Ji Zhou, Lingxuan Meng, and Zhiyong Long. 2020. A practical cross-view image matching method between UAV and satellite for UAV-based geo-localization. *Remote Sensing* 13, 1 (2020), 47.

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

[18] Yunsong Fang, Tingyu Wang, and Zhedong Zheng. 2026. Road Maps as Free Geometric Priors: Weather-Invariant Drone Geo-Localization with GeoFuse. *arXiv preprint arXiv:2605.14925* (2026).

[19] Tongtong Feng, Qing Li, Xin Wang, Mingzi Wang, Guangyao Li, and Wenwu Zhu. 2024. Multi-weather Cross-view Geo-localization Using Denoising Diffusion Models. In *Proceedings of the 2nd Workshop on UAVs in Multimedia: Capturing the World from a New Perspective*. 35–39.

[20] Tongtong Feng, Xin Wang, Feilin Han, Leping Zhang, and Wenwu Zhu. 2024. U2udata: A large-scale cooperative perception dataset for swarm uavs autonomous flight. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 7600–7608.

[21] Yuan Gao, Haibo Liu, and Xiaohui Wei. 2025. Semantic concept perception network with interactive prompting for cross-view image geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology* 35, 6 (2025), 5343–5354.

[22] Fawei Ge, Yunzhou Zhang, Yixiu Liu, Guiyuan Wang, Sonya Coleman, Dermot Kerr, and Li Wang. 2024. Multibranch Joint Representation Learning Based on Information Fusion Strategy for Cross-View Geo-Localization. *IEEE Transactions on Geoscience and Remote Sensing* 62 (2024), 1–16.

[23] Fawei Ge, Yunzhou Zhang, Li Wang, Wei Liu, Yixiu Liu, Sonya Coleman, and Dermot Kerr. 2024. Multi-level Feedback Joint Representation Learning Network Based on Adaptive Area Elimination for Cross-view Geo-localization. *IEEE Transactions on Geoscience and Remote Sensing* (2024).

[24] Payton Goodrich, Omar Betancourt, Ana Claudia Arias, and Tarek Zohdi. 2023. Placement and drone flight path mapping of agricultural soil sensors using machine learning. *Computers and Electronics in Agriculture* 205 (2023), 107591.

[25] Feilin Han, Leping Zhang, Xin Wang, Ke-Ao Zhao, Ying Zhong, Ziyi Su, Tongtong Feng, and Wenwu Zhu. 2024. U2USim-A UAV Telepresence Simulation Platform with Multi-agent Sensing and Dynamic Environment. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 11258–11260.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[27] Qiyi He, Ao Xu, Yifan Zhang, Zhiwei Ye, Wen Zhou, Ruijie Xi, and Qiao Lin. 2024. A Contrastive Learning Based Multiview Scene Matching Method for UAV View Geo-Localization. *Remote Sensing* 16, 16 (2024), 3039.

[28] Shuyu Hu, Zelin Shi, Tong Jin, and Yunpeng Liu. 2025. Query-driven feature learning for cross-view geo-localization. *IEEE Transactions on Geoscience and Remote Sensing* (2025).

[29] Yuxiang Ji, Boyong He, Zhuoyue Tan, and Liaoni Wu. 2024. Game4Loc: A UAV Geo-Localization Benchmark from Game Data. *arXiv preprint arXiv:2409.16925* (2024).

[30] Hao Ju, Shaofei Huang, Si Liu, and Zhedong Zheng. 2025. Video2bev: Transforming drone videos to bevs for video-based geo-localization. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 27073–27083.

[31] Muhammad Arsalan Khan, Wim Ectors, Tom Bellemans, Davy Janssens, and Geert Wets. 2017. UAV-based traffic analysis: A universal guiding framework based on literature survey. *Transportation research procedia* 22 (2017), 541–550.

[32] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. 2017. Learning deep context-aware features over body and latent parts for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 384–393.

[33] Haoran Li, Quan Chen, Zhiwen Yang, and Jiong Yin. 2023. Drone Satellite Matching based on Multi-scale Local Pattern Network. In *Proceedings of the 2023 Workshop on UAVs in Multimedia: Capturing the World from a New Perspective*. 51–55.

[34] Haoran Li, Tingyu Wang, Quan Chen, Qiang Zhao, Shaowei Jiang, Chenggang Yan, and Bolun Zheng. 2024. Aerial-view geo-localization based on multi-layer local pattern cross-attention network. *Applied Intelligence* 54, 21 (2024), 11034–11053.

[35] Haoyuan Li, Chang Xu, Wen Yang, Huai Yu, and Gui-Song Xia. 2024. Learning Cross-view Visual Geo-localization without Ground Truth. *arXiv preprint arXiv:2403.12702* (2024).

[36] Qingge Li, Xiaogang Yang, Jiwei Fan, Ruitao Lu, Bin Tang, Siyu Wang, and Shuang Su. 2024. GeoFormer: An Effective Transformer-based Siamese Network for UAV Geo-localization. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (2024).

[37] Shishen Li, Cuiwei Liu, Huaijun Qiu, and Zhaokui Li. 2023. A transformer-based adaptive semantic aggregation method for UAV visual geo-localization. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer, 465–477.

[38] Yaxuan Li, Jiarui Zeng, Shaofei Huang, and Zhedong Zheng. 2026. Last-Meter Precision Navigation for UAVs: A Diffusion-Refined Aerial Visual Servoing Approach. *arXiv preprint arXiv:2607.04352* (2026).

[39] Jinliang Lin, Zhiming Luo, Dazhen Lin, Shaozi Li, and Zhun Zhong. 2025. A Self-Adaptive Feature Extraction Method for Aerial-View Geo-Localization. *IEEE Transactions on Image Processing* 34 (2025), 126–139.

[40] Jinliang Lin, Zhedong Zheng, Zhun Zhong, Zhiming Luo, Shaozi Li, Yi Yang, and Nicu Sebe. 2022. Joint representation learning and keypoint detection for cross-view geo-localization. *IEEE Transactions on Image Processing* 31 (2022), 3780–3792.

[41] Tsung-Yi Lin, Serge Belongie, and James Hays. 2013. Cross-view image geolocalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 891–898.

[42] Xiangzeng Liu, Ziyao Wang, Yue Wu, and Qiguang Miao. 2024. SeGCN: A Semantic-Aware Graph Convolutional Network for UAV Geo-Localization. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (2024).

[43] David G Lowe. 1999. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, Vol. 2. Ieee, 1150–1157.

[44] Zeng Lu, Tao Pu, Tianshui Chen, and Liang Lin. 2022. Content-aware hierarchical representation selection for cross-view geo-localization. In *Proceedings of the Asian Conference on Computer Vision*. 4211–4224.

[45] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence* 24, 7 (2002), 971–987.

[46] Jie Shao and LingHao Jiang. 2023. Style alignment-based dynamic observation method for UAV-view geo-localization. *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023), 1–14.

[47] Fei Shen, Xiaoyu Du, Liyan Zhang, Xiangbo Shu, and Jinhui Tang. 2023. Triplet contrastive representation learning for unsupervised vehicle re-identification. *arXiv preprint arXiv:2301.09498* (2023).

[48] Fei Shen, Xiangbo Shu, Xiaoyu Du, and Jinhui Tang. 2023. Pedestrian-specific bipartite-aware similarity learning for text-based person retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*. 8922–8931.

[49] Tianrui Shen, Yingmei Wei, Lai Kang, Shanshan Wan, and Yee-Hong Yang. 2023. MCCG: A ConvNeXt-based Multiple-Classifier Method for Cross-view Geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology* (2023).

[50] Jifei Song, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. 2019. Generalizable person re-identification by domain-invariant mapping network. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*. 719–728.

[51] Bo Sun, Ganchao Liu, and Yuan Yuan. 2023. F3-Net: Multi-View Scene Matching for Drone-Based Geo-Localization. *IEEE Transactions on Geoscience and Remote Sensing* (2023).

[52] Haopeng Sun, Yingwei Zhang, Lumin Xu, Sheng Jin, and Yiqiang Chen. 2024. Ultra-high resolution segmentation via boundary-enhanced patch-merging transformer.

[53] Jian Sun, Hao Sun, Lin Lei, Kefeng Ji, and Gangyao Kuang. 2024. TirSA: A Three Stage Approach for UAV-Satellite Cross-View Geo-Localization Based on Self-Supervised Feature Enhancement. *IEEE Transactions on Circuits and Systems for Video Technology* (2024).

[54] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European conference on computer vision (ECCV)*. 480–496.

[55] Xiaoyang Tian, Jie Shao, Deqiang Ouyang, and Heng Tao Shen. 2021. UAV-satellite view synthesis for cross-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 7 (2021), 4804–4815.

[56] Tingyu Wang, Zihao Yang, Quan Chen, Yaoqi Sun, and Chenggang Yan. 2024. Rethinking Pooling for Multi-Granularity Features in Aerial-View Geo-Localization. *IEEE Signal Processing Letters* (2024).

[57] Tingyu Wang, Zhedong Zheng, Yaoqi Sun, Chenggang Yan, Yi Yang, and Tat-Seng Chua. 2024. Multiple-environment Self-adaptive Network for Aerial-view Geo-localization. *Pattern Recognition* 152 (2024), 110363. doi:10.1016/j.patcog.2024.110363

[58] Tingyu Wang, Zhedong Zheng, Chenggang Yan, Jiyong Zhang, Yaoqi Sun, Bolun Zheng, and Yi Yang. 2022. Each Part Matters: Local Patterns Facilitate Cross-View Geo-Localization. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 2 (2022), 867–879. doi:10.1109/TCSVT.2021.3061265

[59] Tingyu Wang, Zhedong Zheng, Zunjie Zhu, Yaoqi Sun, Chenggang Yan, and Yi Yang. 2024. Learning Cross-View Geo-Localization Embeddings via Dynamic Weighted Decorrelation Regularization. *IEEE Transactions on Geoscience and Remote Sensing* 62 (2024), 1–12. doi:10.1109/TGRS.2024.3491757

[60] Yuntao Wang, Jinpu Zhang, Ruonan Wei, Wenbo Gao, and Yuehuan Wang. 2024. Mfrgn: Multi-scale feature representation generalization network for ground-to-aerial geo-localization. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 2574–2583.

[61] Zhen Wang, Dianxi Shi, Chunping Qiu, Songchang Jin, Tongyue Li, Yanyan Shi, Zhe Liu, and Ziteng Qiao. 2024. Sequence matching for Image-Based UAV-to-Satellite Geolocalization. *IEEE Transactions on Geoscience and Remote Sensing* (2024).

[62] Qiong Wu, Yi Wan, Zhi Zheng, Yongjun Zhang, Guangshuai Wang, and Zhenyang Zhao. 2024. CAMP: A Cross-View Geo-Localization Method Using Contrastive Attributes Mining and Position-Aware Partitioning. *IEEE Transactions on Geoscience and Remote Sensing* 62 (2024), 1–14. doi:10.1109/TGRS.2024.3448499

[63] Panwang Xia, Yi Wan, Zhi Zheng, Yongjun Zhang, and Jiwei Deng. 2024. Enhancing Cross-View Geo-Localization With Domain Alignment and Scene Consistency. *IEEE Transactions on Circuits and Systems for Video Technology* 34, 12 (2024), 13271–13281. doi:10.1109/TCSVT.2024.3443510

[64] Linhui Xiao, Xiaoshan Yang, Fang Peng, Yaowei Wang, and Changsheng Xu. 2024. Hivg: Hierarchical multimodal fine-grained modulation for visual grounding. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 5460–5469.

[65] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. 2018. Attention-aware compositional network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2119–2128.

[66] Mo Yang, Luo Chen, and Ning Jing. 2026. Integrating Multiscale Consistency and Enhanced Feature Interaction for Cross-View Geo-Localization. *IEEE Transactions on Geoscience and Remote Sensing* 64 (2026), 1–16.

[67] Qin Ye, Junqi Luo, and Yi Lin. 2024. A coarse-to-fine visual geo-localization method for GNSS-denied UAV with oblique-view imagery. *ISPRS Journal of Photogrammetry and Remote Sensing* 212 (2024), 306–322.

[68] Xu Yin, Woobin Im, Dongbo Min, Yuchi Huo, Fei Pan, and Sung-Eui Yoon. 2024. Fine-grained Background Representation for Weakly Supervised Semantic Segmentation. *IEEE Transactions on Circuits and Systems for Video Technology* (2024).

[69] Haotian Zhang, Gaoang Wang, Zhichao Lei, and Jenq-Neng Hwang. 2019. Eye in the sky: Drone-based object tracking and 3d localization. In *Proceedings of the 27th ACM International Conference on Multimedia*. 899–907.

[70] Hu Zhao, Keyan Ren, Tianyi Yue, Chun Zhang, and Shuai Yuan. 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* (2024).

[71] Chenhao Zheng, Ayush Shrivastava, and Andrew Owens. 2023. Exif as language: Learning cross-modal associations between images and camera metadata. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6945–6956.

[72] Chenhao Zheng, Jieyu Zhang, Aniruddha Kembhavi, and Ranjay Krishna. 2024. Iterated learning improves compositionality in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13785–13795.

[73] Zhedong Zheng, Xiaohan Wang, Nenggan Zheng, and Yi Yang. 2022. Parameter-efficient person re-identification in the 3d space. *IEEE Transactions on Neural Networks and Learning Systems* (2022).

[74] Zhedong Zheng, Yunchao Wei, and Yi Yang. 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.

[75] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. 2019. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 598–607.

[76] Runzhe Zhu, Ling Yin, Mingze Yang, Fei Wu, Yuncheng Yang, and Wenbo Hu. 2023. SUES-200: A Multi-height Multi-scene Cross-view Image Benchmark Across Drone and Satellite. *IEEE Transactions on Circuits and Systems for Video Technology* (2023).

[77] Sijie Zhu, Mubarak Shah, and Chen Chen. 2022. Transgeo: Transformer is all you need for cross-view image geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1162–1171.

[78] Jiedong Zhuang, Ming Dai, Xuruoyan Chen, and Enhui Zheng. 2021. A faster and more effective cross-view matching method of uav and satellite images for uav geolocalization. *Remote Sensing* 13, 19 (2021), 3979.

[79] F Zimmermann, C Eling, L Klingbeil, and H Kuhlmann. 2017. Precise positioning of uavs–dealing with challenging rtk-gps measurement conditions during automated uav flights. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 4 (2017), 95–102.