From Coarse to Fine: A Matching and Alignment Framework for Unsupervised Cross-View Geo-Localization

Xueyi Wang[1], Lele Zhang[1], Zheng Fan[1], Yang Liu[1], Chen Chen[1], Fang Deng[1,2*]

[1]Beijing Institute of Technology
[2]Beijing Institute of Technology Chongqing Innovation Center
{wangxueyi, zhanglele, fanzheng, liuyang, xiaofan, dengfang}@bit.edu.cn

Abstract

Cross-view geo-localization aims at determining the geographic location of a query image by matching the reference images. The matching pairs can be captured from diverse perspectives, such as those from satellites and drones. Most existing methods are supervised that require input of location-labeled images or matched and unmatched image pairs for training, resulting in high labor costs. Moreover, current unsupervised methods perform instances matching directly between different perspectives with dramatic discrepancies, resulting in poor performance. To address these issues, this paper proposes a novel matching and alignment framework from coarse instance-cluster level to fine intermediate instance level for unsupervised cross-view geo-localization. We first introduces cluster-based contrastive learning, assigning pseudo-labels to the instances and generate clusters within each view. Then we design a cross-view location alignment module that fully exploits the feature relationships between instances and clusters for intra- and inter-views. Finally, we design an intermediate state transition module that facilitates further alignment between views by constructing intermediate states and bringing both views closer to the intermediate domain simultaneously. Extensive experiments demonstrate that our method surpasses state-of-the-art unsupervised cross-view geo-localization methods and even achieves comparable performance to state-of-the-art supervised methods.

Introduction

Cross-view geo-localization aims to locate an image by retrieving images from the same location but different views in a large-scale gallery (Workman, Souvenir, and Jacobs 2015). Different from traditional vision-based geometry geo-localization (Zhang et al. 2018; Deng et al. 2020; Gao et al. 2021; Zhang et al. 2022), this research is capable of autonomous geo-localization in the absence or weakness of the global navigation satellite system (GNSS) signal, with more robust and flexible localization capabilities that enable widespread appliances across various domains, including unmanned ground vehicles and drone localization and navigation in GNSS-denied environments (Shetty and Gao 2019; Shi et al. 2020a). Images captured from ground and satellite perspectives are frequently utilized in cross-view geo-localization (Liu and Li 2019; Shi et al. 2020b, 2019; Sun et al. 2019). However, cross-view geo-localization remains a major challenge due to the orthogonality of the two views, resulting in limited feature intersections. The advancement of drone technology has introduced drone view to enhance cross-view geo-localization(Zheng, Wei, and Yang 2020; Lin et al. 2022; Sun, Liu, and Yuan 2023; Wang et al. 2022). The drone view plays a transitional role in image matching between the ground and satellite views, and images collected from the drone have more features overlapping with the images from the other two views. Therefore, drone-satellite geo-localization is proposed, including drone-view target localization and drone navigation (Zheng, Wei, and Yang 2020). The former aims to locate the target by retrieving the corresponding satellite image. The latter is to guide the drone to its destination by matching the given satellite image with the drone image located at the same location in the gallery.

Recently, more and more cross-view geo-localization datasets with labels have been proposed (Zhu, Yang, and Chen 2021; Zheng, Wei, and Yang 2020). Leveraging these datasets, cross-view geo-localization has achieved excellent effectiveness along with the evolution of deep learning. Existing optimization objectives for cross-view geo-localization methods can be broadly categorized into two main groups: one aims to distinguish between matched and unmatched pairs by pulling matched pairs closer and pushing unmatched pairs apart (Hu et al. 2018; Liu and Li 2019), and the other is to define the problem as a classification task, which refers to learning discriminative representation for different location classes (Wang et al. 2021; Shen et al. 2023). All these methods have requirements on the input, either images with known locations or matched and unmatched image pairs, which significantly consumes time and human labor. In addition, not all images in reality come with matched images in another view or location tags. Therefore, we introduce the unsupervised method for cross-view geo-localization. Our objective is to learn view-invariant features and distinguish different location classes to achieve cross-view matching in cases where the image labels are unknown and no matching image pairs can be found.

For unsupervised cross-view geo-localization, there are two main challenges in bridging the gap between views and performing intra- and inter-view location associations of im-

*Corresponding author

ages without any annotation. Existing unsupervised cross-view geo-localization methods assign pseudo-labels and connect different perspectives through generating pseudo-samples (Li, Qian, and Xia 2024) or employing similarity between samples (Li et al. 2024), thus training and updating the network in a supervised-like manner. The above steps are performed at the instance level, which lacks the consideration of global information and leads the network to focus only on the selected individuals in a single iteration, susceptible to noisy samples. Summarizing the above analysis, we adopt a cluster-then-refine pattern (Dai et al. 2022) as our fundamental methodology, which assigns pseudo-labels to samples with high similarity within the same perspective using a clustering algorithm. To establish label associations and bridge the gap between different perspectives, we propose a cross-view location alignment module that performs efficient cluster matching through a local-to-global cluster matching strategy and leverages both instances and cluster centroids to align the same categories across different views. In addition, to decrease the significant feature discrepancy between perspectives, we propose an intermediate state transition module that explicitly decreases the color disparity between satellite-view and drone-view images and constructs intermediate domains to which both perspectives approach at the same time.

The primary contributions of our work are outlined as follows:

- We propose a novel framework for unsupervised cross-view geo-localization without data annotation. We introduce cluster-based contrastive learning to assign pseudo-labels to samples and form clusters inside views.
- We develop a novel cross-view location alignment module that forms pseudo-label connections across views through a local-to-global cluster matching strategy. By leveraging instances and clustering centroids, the proposed module enables reliable feature alignment between views and reduces susceptibility to noisy samples.
- We design an intermediate state transition module. A non-learning method is introduced to construct intermediate states in the absence of matched pairs and two view features are simultaneously aligned to achieve finer alignment.
- We validate the efficacy of our proposed method on University-1652 and SUES-200 datasets. Extensive experiments reveal that our method outperforms the state-of-the-art unsupervised approaches and even achieves comparable performance to state-of-the-art supervised approaches.

Related Work

Cross-View Geo-Localization

In the early days, some hand-crafted features were used for cross-view geo-localization (Lin, Belongie, and Hays 2013; Bansal et al. 2011; Castaldo et al. 2015; Senlet and Elgammal 2011). Owing to their poor robustness and difficulty in extracting valuable features amidst drastically changing views, deep features are adopted, accompanied by the evolution of deep learning. CNN (Simonyan and Zisserman 2014; He et al. 2016) and Transformer (Vaswani et al. 2017; Zheng et al. 2024) are both widely used frameworks for cross-view geo-localization. Wang et al. (Wang et al. 2021) presented a square-ring partitioning strategy to exploit the target building and its surrounding environment, thus effectively learning contextual information. Yang et al. (Yang, Lu, and Zhu 2021) proposed a hybrid approach using a Transformer on CNN. Zhu et al. (Zhu, Shah, and Chen 2022) and Dai et al. (Dai et al. 2021) introduced pure transformer frameworks for feature extraction. All of these methods require inputting matched and unmatched image pairs, or images with known location categories, into the model, which is highly labor-intensive.

Li et al. (Li, Qian, and Xia 2024) proposed an unsupervised cross-view geo-localization method for satellite and ground views using generated pseudo-satellite images to correlate the two views. Li et al. (Li et al. 2024) introduced an expectation maximization algorithm to assign pseudo-labels to unlabeled samples. The methods mentioned above bridge the domain gap and learn perspective-invariant feature representations at the instance level, losing global information about the samples to some extent, which may lead to an accuracy decrease under the influence of noisy samples. Therefore, we introduce clustering-based contrastive learning for unsupervised cross-view geo-localization, aiming at acquiring cross-view feature representations at the instance level and the centroid level, without relying on labeled data.

Unsupervised Contrastive Learning

For unsupervised contrastive learning, the primary challenge lies in constructing positive and negative samples when labeled data is unavailable (Zhu et al. 2023a). A common way is to use augmented query samples as positive samples and other samples as negative samples (Wu et al. 2018; Zhang, Lin, and Liu 2023; Zhu et al. 2024). To ensure an ample supply of negative samples for comparison, He et al. (He et al. 2020) stored the samples through a dynamic queue. Chen et al. (Chen et al. 2020) set a huge batch size and treated the samples within the batch as negative samples for each other. Limited by computational resources, these methods lack comparison with global information. Caron et al. (Caron et al. 2020) proposed an online clustering method that aligned samples to cluster centers and allowed comparison with global features without requiring significant computational resources. Dai et al. (Dai et al. 2022) proposed ClusterNCE, which regarded cluster centroids as positive and negative samples and calculated loss at the cluster level.

Method

This section introduces our proposed unsupervised method, depicted in Fig. 1. First, we describe the problem definition. Then, we elaborate the important components of our proposed method, including the baseline based on cluster contrastive learning, the cross-view location alignment module and the intermediate state transition module.

Problem fomulation. Given a cross-view geo-localization dataset $\{X^s, X^d\}$, $X^s = \{x_i^s | i = 1, 2, ..., N^s\}$

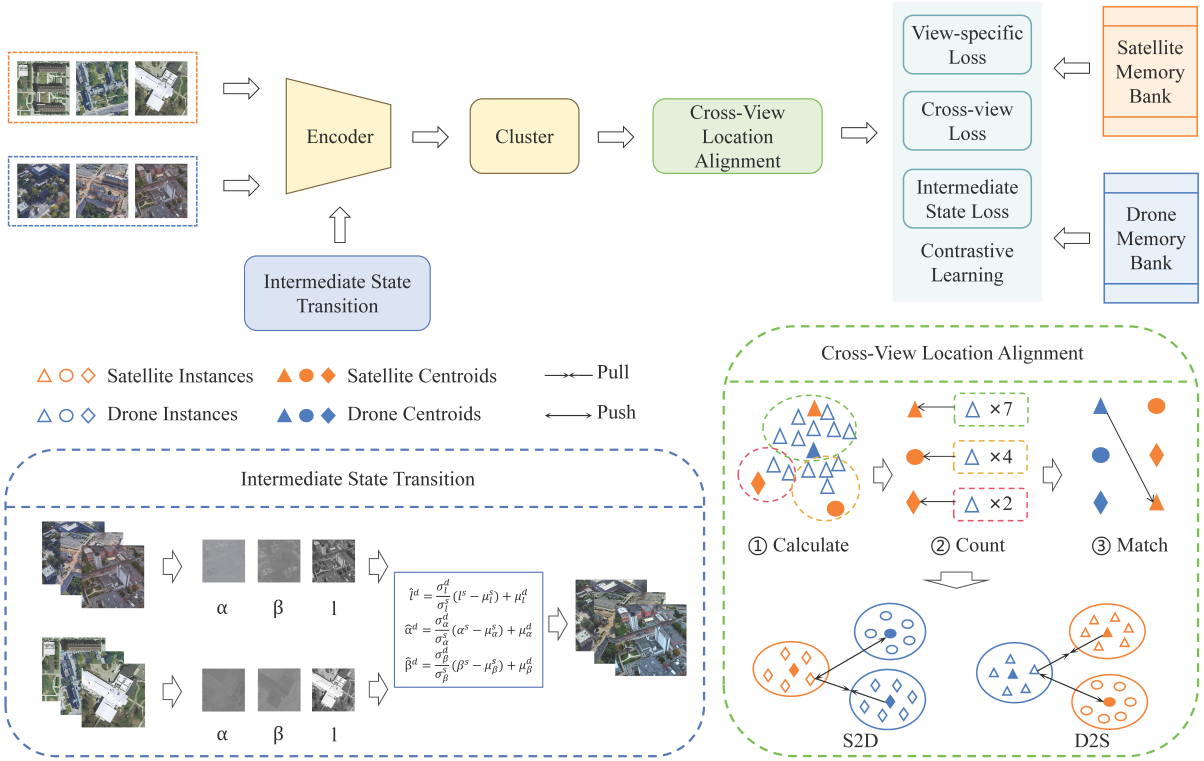

Figure 1: The structure of our proposed method. At the initialization of each epoch, feature extraction and clustering are conducted separately on two views. Subsequently, pseudo-labels are allocated to each instance according to the clustering results, and cluster centroids are computed to construct memory banks. The cross-view location alignment module performs matching and implements cluster-instance alignment across views. The intermediate state transition module generates intermediate states for the transition of two views to the intermediate domain, achieving a finer alignment.

represents the input images in satellite view, and $X^d = \{x_j^d | j = 1, 2, ..., N^d\}$ represents the input images in drone view. N^s and N^d represent the number of images in two views, respectively. Images in both views are without any labels. The objective of unsupervised cross-view geo-localization is to learn view-invariant and location-related feature representation without annotations. With the proposed model, when the query is a satellite image, the corresponding image is retrieved in the drone-view image dataset and vice versa.

Cluster Contrastive Learning Baseline

Inspired by (Dai et al. 2022), we introduce cluster contrastive learning for unsupervised cross-view geo-localization, in order to utilize the identifiable characteristics of individual samples and the global characteristics of the overall data.

Memory Initialization The memory banks are initialized at the start of each training epoch. First, the DBSCAN algorithm (Ester et al. 1996) is utilized to cluster the feature vectors obtained from the feature extractors and assign pseudo labels to the feature vectors with the following formulation:

$$\hat{Y^s} = \mathbf{DBSCAN}(F(X^s; \theta^s)) \tag{1}$$

$$\hat{Y^d} = \mathbf{DBSCAN}(F(X^d; \theta^d)) \tag{2}$$

where $\mathbf{DBSCAN}(\cdot)$ denotes the clustering algorithm, $F(\cdot; \theta^s)$ and $F(\cdot; \theta^d)$ denote the satellite and drone feature extractors. $\hat{Y^s} = \{\hat{y}_1^s, \hat{y}_2^s, \ldots, \hat{y}_{N^s}^s\}$ and $\hat{Y^d} = \{\hat{y}_1^d, \hat{y}_2^d, \ldots, \hat{y}_{N^d}^d\}$ denote the pseudo labels of the input images X^s and X^d.

Then, we calculate the cluster centroids to initialize the memory banks:

$$c_k^s = \frac{1}{|\mathcal{C}_k^s|} \sum_{f_i^s \in \mathcal{C}_k^s} f_i^s \quad (k = 1, 2, \ldots, N_c^s) \tag{3}$$

$$c_l^d = \frac{1}{|\mathcal{C}_l^d|} \sum_{f_j^d \in \mathcal{C}_l^d} f_j^d \quad (l = 1, 2, \ldots, N_c^d) \tag{4}$$

where c_k^s and c_l^d represent the kth and lth cluster centroids stored in the satellite and drone memory banks. $\mathcal{C}_k^s$ and $\mathcal{C}_l^d$ represent the kth satellite-view cluster set and the lth drone-view cluster set. N_c^s and N_c^d represent the number of satellite-view and drone-view categories obtained with the clustering algorithm. $|\cdot|$ denotes the number of feature vectors in a cluster set. f_i^s and f_j^d denotes the feature vectors extracted by the feature extractors $F(\cdot; \theta^s)$ and $F(\cdot; \theta^d)$, respectively.

Memory Updating During a single epoch in training, the momentum update mechanism (He et al. 2020) are employed to perform updating and maintain partial consistency in different mini-batches of an epoch. We perform a momentum update to the memory banks using instance feature vectors in each iteration:

$$c_k^s[t] \leftarrow mc_k^s[t-1] + (1-m)q_{ki}^s \tag{5}$$

$$c_l^d[t] \leftarrow mc_l^d[t-1] + (1-m)q_{li}^d \tag{6}$$

where q_{ki}^s and q_{li}^d denote the ith query feature vector of the kth satellite and lth drone cluster, t denotes the tth iteration in an epoch, and m denotes the momentum. m indicates the proportion of the updated cluster centroids between the former cluster centroids and the latest query feature vectors.

View-Specific Cluster Contrastive Learning We independently employ ClusterNCE loss (Dai et al. 2022) for the two branches to distinguish the instances at different locations within the views and to bring the instances closer to the centroid to which they belong. The loss function of view-specific cluster contrastive learning is divided into two parts:

$$L_{SCC} = -log\frac{exp(q^s \cdot c_+^s/\tau)}{\sum_{k=0}^{N_c^s} exp(q^s \cdot c_k^s/\tau)} \tag{7}$$

$$L_{DCC} = -log\frac{exp(q^d \cdot c_+^d/\tau)}{\sum_{l=0}^{N_c^d} exp(q^d \cdot c_l^d/\tau)} \tag{8}$$

where τ represents the temperature hyper-parameter, c_+^s and c_+^d are the positive cluster centroids that the query instances q^s and q^d belong to. The total loss function of view-specific cluster contrastive learning is formulated by:

$$L_{VCC} = L_{SCC} + L_{DCC} \tag{9}$$

Cross-View Location Alignment

To mitigate the gap and establish the bridge between views, we introduce the cross-view location alignment module, which involves local-to-global cross-view cluster matching and cross-view cluster contrastive learning.

Local-to-Global Cross-View Cluster Matching Cross-view cluster matching aims at matching clusters from one view with the corresponding clusters from another, with the matched clusters considered to be from the same location. To perform reliable matching, we propose a new local-to-global cross-view cluster matching strategy. The strategy employs interactive matching from local instances to global cluster centroids instead of centroid-only or instance-only, reducing the impact of sample noise on cluster matching.

Taking the matching from drone view to satellite view as an example, we first compute the cosine similarity between each drone instance feature vector and the satellite cluster centroid:

$$sim(f_i^d, c_k^s) = \frac{f_i^d \cdot c_k^s}{\|f_i^d\| \times \|c_k^s\|} \tag{10}$$

We designate the satellite-view cluster centroid with the highest similarity as the corresponding cluster centroid for the given drone-view instance:

$$k_i^* = arg\max_k sim(f_i^d, c_k^s) \ \ (k = 1, 2, \ldots, N_c^s) \tag{11}$$

where k_i^* indicates the index of satellite-view cluster centroid that matches the ith drone-view instance.

Following Eq.(11), we count the instances within lth drone-view cluster that belong to each satellite-view centroid:

$$count(\mathcal{C}_l^d, \mathcal{C}_k^s) = \sum_{f_i^d \in C_l^d} \delta(f_i^d, c_k^s) \ \ (k = 1, 2, \ldots, N_c^s) \tag{12}$$

where $\delta()$ is given as follows:

$$\delta(f_i^d, c_k^s) = \begin{cases} 1, \text{if } k_i^* = k \\ 0, \text{otherwise.} \end{cases} \tag{13}$$

Finally, we establish a match from the lth drone cluster to the $k^{**}th$ satellite cluster if and only if the count of instances within the lth drone cluster, aligning with the corresponding $k^{**}th$ satellite-view cluster centroid, is maximized. It is expressed as:

$$match(l, k^{**}) = \begin{cases} 1, \text{if } k^{**} = arg\max_k count(\mathcal{C}_l^d, \mathcal{C}_k^s) \\ 0, \text{otherwise.} \end{cases} \tag{14}$$

The matching process from satellite view to drone view is the same as the process described above. The distinction lies in matching satellite instances to the drone clustering centroids. Our proposed strategy ensures that the matching result from satellite to drone is independent of that from drone to satellite, which improves the fault tolerance of the matching.

Cross-view Cluster Contrastive Learning Building on cross-view cluster matching, cross-view cluster contrastive learning is utilized to learn view-invariant location-specific feature representations across views. The loss function takes the following form:

$$L_{S2D} = -log\frac{exp(q^s \cdot c_+^d/\tau)}{\sum_{l=0}^{N_c^d} exp(q^s \cdot c_l^d/\tau)} \tag{15}$$

$$L_{D2S} = -log\frac{exp(q^d \cdot c_+^s/\tau)}{\sum_{k=0}^{N_c^s} exp(q^d \cdot c_k^s/\tau)} \tag{16}$$

where c_+^s is the satellite-view cluster centroid matching the drone-view query instance q^d and c_+^d is the drone-view cluster centroid matching the satellite-view query instance q^s. The cross-view cluster contrastive learning will pull closer between query instances and the corresponding cross-view clustering centroids, thereby bridging the cross-view gap.

The total loss function of cross-view location alignment is expressed as:

$$L_{CVCC} = L_{S2D} + L_{D2S} \tag{17}$$

Intermediate State Transition

Direct cross-view alignment is challenging due to the significant discrepancies between the two views. Therefore, we propose the intermediate state transition module to construct an intermediate domain in which features of the two views are gathered to align.

Intermediate State Construction Images in two views exhibit variations in color at the same location due to differences in data sources, capture equipment, time, weather, etc. However, images at the same location still have relatively consistent color distributions, which our eyes can use to determine whether they are from the same location. Inspired by (Reinhard et al. 2001), we propose a non-learning intermediate state generation method without the need of matched image pairs.

In RGB space, the pixel values of the three channels interact with each other and are not independent. To perform effective color transfer, we initially transform the images from the RGB space to the $l\alpha\beta$ space with less correlation between the axes. Due to the high imaging quality in drone view and the limited number of satellite-view images, we construct intermediate states by performing color transfer to drone-view images to enhance the effectiveness of the intermediate state transition module. The values of the three axes in the source drone images and satellite images are denoted as $(l^d, \alpha^d, \beta^d), (l^s, \alpha^s, \beta^s)$, respectively. Then, we calculate the mean value $(\mu_l^d, \mu_\alpha^d, \mu_\beta^d), (\mu_l^s, \mu_\alpha^s, \mu_\beta^s)$ and standard deviation $(\sigma_l^d, \sigma_\alpha^d, \sigma_\beta^d), (\sigma_l^s, \sigma_\alpha^s, \sigma_\beta^s)$ on each axis of the images. The target intermediate state is calculated as follows:

$$\hat{l}^d = \frac{\sigma_l^d}{\sigma_l^s}(l^s - \mu_l^s) + \mu_l^d \tag{18}$$

$$\hat{\alpha}^d = \frac{\sigma_\alpha^d}{\sigma_\alpha^s}(\alpha^s - \mu_\alpha^s) + \mu_\alpha^d \tag{19}$$

$$\hat{\beta}^d = \frac{\sigma_\beta^d}{\sigma_\beta^s}(\beta^s - \mu_\beta^s) + \mu_\beta^d \tag{20}$$

Due to the absence of labels in unsupervised learning, we use a global computation strategy to improve the color transfer module. We firstly obtain the mean and standard deviation of each axis for all images in the satellite-view dataset beforehand and then average them again as parameters to perform color transfer on the drone-view images:

$$\mu_l^s = \frac{\sum_{i=1}^{N^s} \mu_{li}^s}{N^s},\ \sigma_l^s = \frac{\sum_{i=1}^{N^s} \sigma_{li}^s}{N^s} \tag{21}$$

$$\mu_\alpha^s = \frac{\sum_{i=1}^{N^s} \mu_{\alpha i}^s}{N^s},\ \sigma_\alpha^s = \frac{\sum_{i=1}^{N^s} \sigma_{\alpha i}^s}{N^s} \tag{22}$$

$$\mu_\beta^s = \frac{\sum_{i=1}^{N^s} \mu_{\beta i}^s}{N^s},\ \sigma_\beta^s = \frac{\sum_{i=1}^{N^s} \sigma_{\beta i}^s}{N^s} \tag{23}$$

Finally, we get each axis value of the target intermediate state by the Eq. (18)-Eq. (20) and convert it from $l\alpha\beta$ space to RGB space to realize the unsupervised color transfer. Examples of intermediate states are shown in supplementary material.

Intermediate State-based Cluster Contrastive Learning We realize the gathering of the instances in two views to the intermediate domain by intermediate state-based cluster contrastive learning, narrowing the discrepancy between views with the aid of the intermediate domain. The functions are expressed as:

$$L_{I2S} = -log\frac{exp(q^i \cdot c_+^s/\tau)}{\sum_{k=0}^{N_c^s} exp(q^i \cdot c_k^s/\tau)} \tag{24}$$

$$L_{I2D} = -log\frac{exp(q^i \cdot c_+^d/\tau)}{\sum_{l=0}^{N_c^d} exp(q^i \cdot c_l^d/\tau)} \tag{25}$$

The total loss function of intermediate state-based cluster contrastive learning is expressed as:

$$L_{ISCC} = L_{I2S} + L_{I2D} \tag{26}$$

Optimization Objective

Finally, the above loss functions of the view-specific, cross-view, and intermediate state-based contrastive learning are integrated into one framework, enabling intra- and inter-view sample alignment from multiple levels. The model is trained using the following loss function:

$$L = L_{VCC} + L_{ISCC} + L_{CVCC} \tag{27}$$

Experiments

Experimental Setup

Datasets and Evaluation Metric Our proposed method focuses on cross-view matching between drone and satellite perspectives, including drone-view target localization (drone to satellite) and drone navigation (satellite to drone). The evaluation is conducted on two widely-used cross-view geo-localization datasets, University-1652 (Zheng, Wei, and Yang 2020), and SUES-200 (Zhu et al. 2023b).

University-1652 (Zheng, Wei, and Yang 2020) is a large-scale benchmark dataset gathered from 1,652 buildings at 72 universities worldwide. The training set includes 50,218 images from 701 buildings at 33 universities, with 37,854 drone images and 701 satellite images. The testing set is from 951 buildings at 39 universities, ensuring no overlap between the training and testing sets.

SUES-200 (Zhu et al. 2023b) is a multi-scene, multi-height, cross-view geo-localization dataset that gathers drone and satellite images from 200 locations. Specifically, the drone-view images are captured from four altitude, i.e. 150, 200, 250, 300m. Only one altitude is used in a complete training and testing cycle. The dataset has 10000 drone images per altitude and 200 satellite images, of which 60% are allocated to the training set and 40% to the query set. The gallery set is the sum of training and query sets, with the training set serving as distractors.

The experimental evaluation utilizes Recall@K (R@K) and average precision (AP) metrics to assess the efficacy of our proposed method. Recall@K denotes the probability that the true-matched images appear in the top-K of the ranking list. Higher Recall@K values signify greater retrieval accuracy. AP represents the area under the precision-recall (PR) curve, reflecting the positions of all true-matched images within the ranking list.

Method	Type	Drone to Satellite		Satellite to Drone	
		R@1	AP	R@1	AP
Zheng et al. (Zheng, Wei, and Yang 2020)	Supervised	59.69	64.8	73.18	59.4
LCM (Ding et al. 2020)	Supervised	66.65	70.82	79.89	65.38
LPN (Wang et al. 2021)	Supervised	75.93	79.14	86.45	74.79
PCL (Tian et al. 2021)	Supervised	79.47	83.63	87.69	78.51
F3-net (Sun, Liu, and Yuan 2023)	Supervised	78.64	81.60	-	-
FSRA (Dai et al. 2021)	Supervised	82.25	84.82	87.87	81.53
MCCG (Shen et al. 2023)	Supervised	**89.28**	**91.01**	**94.29**	**89.29**
Li et al. (Li et al. 2024)	Unsupervised	70.29	74.93	79.03	61.03
Ours	Unsupervised	**85.95**	**90.33**	**94.01**	**82.66**

Table 1: Comparison results with state-of-the-art supervised cross-view geo-localization methods and state-of-the-art unsupervised methods in University-1652.

Drone to Satellite									
Method	Type	150m		200m		250m		300m	
		R@1	AP	R@1	AP	R@1	AP	R@1	AP
SUES-200 Baseline (Zhu et al. 2023b)	Supervised	55.65	61.92	66.78	71.55	72.00	76.43	74.05	78.26
ViT (Zhu et al. 2023b)	Supervised	59.32	64.94	62.30	67.22	71.35	75.48	77.17	80.67
LCM (Ding et al. 2020)	Supervised	43.42	49.65	49.42	55.91	54.47	60.31	60.43	65.78
LPN (Wang et al. 2021)	Supervised	61.58	67.23	70.85	75.96	80.38	83.80	81.47	84.53
FSRA (Dai et al. 2021)	Supervised	68.25	73.45	83.00	85.99	90.68	92.27	91.95	93.46
MCCG (Shen et al. 2023)	Supervised	**82.22**	**85.47**	**89.38**	**91.41**	**93.82**	**95.04**	**95.07**	**96.20**
Ours	Unsupervised	**76.90**	**84.95**	**87.88**	**92.60**	**92.98**	**95.66**	**95.10**	**96.92**
Satellite to Drone									
Method	Type	150m		200m		250m		300m	
		R@1	AP	R@1	AP	R@1	AP	R@1	AP
SUES-200 Baseline (Zhu et al. 2023b)	Supervised	75.00	55.46	85.00	66.05	86.25	69.94	88.75	74.46
ViT (Zhu et al. 2023b)	Supervised	82.50	58.88	87.50	62.48	90.00	69.91	96.25	84.10
LCM (Ding et al. 2020)	Supervised	57.50	38.11	68.75	49.19	72.50	47.94	75.00	59.36
LPN (Wang et al. 2021)	Supervised	83.75	66.78	88.75	75.01	92.50	81.34	92.50	85.72
FSRA (Dai et al. 2021)	Supervised	83.75	76.67	90.00	85.34	93.75	90.17	95.00	92.03
MCCG (Shen et al. 2023)	Supervised	**93.75**	**89.72**	**93.75**	**92.21**	**96.25**	**96.14**	**98.75**	**96.64**
Ours	Unsupervised	**87.50**	**74.81**	**92.50**	**87.15**	**96.25**	**91.20**	**98.75**	**94.52**

Table 2: Comparative results with state-of-the-art supervised cross-view geo-localization methods and state-of-the-art unsupervised methods in SUES-200.

Implementation details The encoder utilizes AGW (Ye et al. 2021), consisting of a shallow network without weight sharing and a weight-shared ResNet50 (He et al. 2016) whose weights are pre-trained on ImageNet. The feature vectors produced through the encoder are 2048 dimensional for subsequent learning and testing. We perform basic image augmentation, such as flipping and random cropping on satellite images before training to simulate different scenarios and enrich satellite samples for effective clustering. The images are resized to 224×224 pixels throughout our experiments. The batch size is configured to 128. We utilize the Adam optimizer with a weight decay of 5e-4 and an initial learning rate of 3.5e-3, which is reduced by 0.1 every ten epochs. The maximum distance of DBSCAN is 0.3. The temperature factor τ is 0.05, and the momentum m is 0.1. The model is trained for 50 epochs in total. In testing, we employ the cosine distance to quantify the similarity between the query sample and the candidates in the gallery set. All experiments are performed on Nvidia RTX 3090 GPUs.

Comparison with State-of-the-arts

Comparison with the Supervised Methods We compare the performance of our proposed unsupervised cross-view geo-localization method with the state-of-the-art supervised methods on University-1652 and SUES-200. The comparison results are shown in Table 1 and Table 2. Our method achieves comparable results to the state-of-the-art supervised methods, and even obtains the highest scores on some metrics. These results show that our proposed unsupervised method can effectively perform cross-view geo-localization with robustness at different heights in the absence of labels. However, our results are slightly below the state-of-the-art

BL	CLA	IST	Drone to Satellite R@1	Drone to Satellite AP	Satellite to Drone R@1	Satellite to Drone AP
✓			59.10	68.63	72.61	49.60
✓	✓		84.55	89.33	93.01	80.04
✓		✓	80.15	86.10	91.44	76.62
✓	✓	✓	**85.95**	**90.33**	**94.01**	**82.66**

Table 3: Ablation studies in University-1652. BL: baseline.

Method	Drone to Satellite R@1	Drone to Satellite AP	Satellite to Drone R@1	Satellite to Drone AP
Instance-only	79.35	85.30	85.59	66.09
Cluster-only	81.73	87.24	92.30	78.98
Ours	**84.55**	**89.33**	**93.01**	**80.04**

Table 4: Comparative results with the state-of-the-art cluster matching method in University-1652.

supervised methods on other metrics. We think this may be caused by errors in clustering and matching and data discrepancy.

Comparison with Unsupervised Methods To evaluate the performance of our method compared to unsupervised methods, we conduct comparative experiments against the currently only unsupervised satellite-drone geo-localization method on University-1652. However, due to the unavailability of its source code, we are unable to compare its results on the SUES-200. As shown in Table 1, our method outperforms the state-of-the-art unsupervised method in drone-to-satellite and satellite-to-drone. Compared to SOTA, our proposed method obtains about 15%-20% improvement in all evaluation metrics.

Ablation Studies

We carry out several ablation experiments to demonstrate the efficacy of our method.

Effectiveness of Cross-View Location Alignment The purpose of the cross-view location alignment module is to perform category matching between views in the absence of real labels, thereby learning view-invariant location-specific features to establish associations for samples at the same location and to have discrimination for samples at different locations across views. We verify the effectiveness of the cross-view location alignment block by adding it to the baseline. As shown in Table 3, we know that the cross-view location alignment module bridges two views and facilitates the model to realize location alignment across views.

Effectiveness of Intermediate State Transition We verify the validity of the intermediate state transition module by adding it to the baseline alone and with the cross-view location alignment module. As shown in Table 3, the results demonstrate that the intermediate state transition block promotes the model's capability to align the same location in different views. It brings the features of both views closer to the intermediate transition domain at the same time, where cross-view contrastive learning is more effective than the direct one.

Effectiveness of Local-to-global Cross-View Cluster Matching To validate the effectiveness of our proposed cross-view cluster matching strategy, we compare it with instance-only and cluster-only matching methods, respectively. For a fairer comparison, we remove the intermediate state transition module in our model. As can be seen from Table 4, our proposed local-to-global cross-view cluster matching strategy considers both instances and cluster centroids, and the matching results from drone to satellite and from satellite to drone are not identical in one matching round, which can correct the clustering error and reduce the negative impact of noisy samples to some extent.

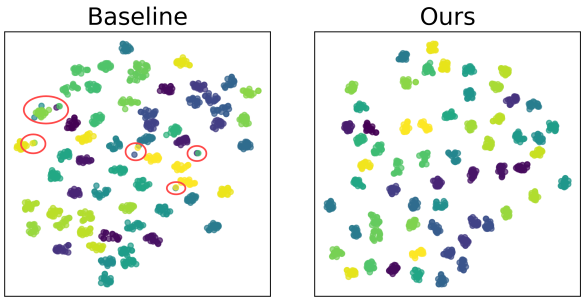

Figure 2: The t-SNE visualization of baseline and our proposed method. There are 60 location categories randomly selected, each color representing a location. (a) Baseline. (b) Ours. Red circles represent false matches.

Visualization of Qualitative Results

We perform visualization of the feature vectors extracted by baseline and our method through t-SNE map. As shown in Fig. 2, for baseline, the drone-view feature vectors of the same category are pulled closer, and those of different categories are pushed away. However, the drone-view feature vectors are not aggregated with the satellite-view feature vectors of the same category. With our model, the feature vectors of two views belonging to the same category are pulled closer together, and those belonging to different categories are pushed farther apart. Eventually, all samples of the same category are gathered.

Conclusion

In this work, we argue that labels are dispensable and present a novel unsupervised cross-view geo-localization framework. Unlike already existing unsupervised cross-view geo-localization methods, we propose a new cross-view location alignment module to establish reliable connections between views, fully leveraging individual instances and global clusters. In addition, we propose the intermediate state transition module to bridge the significant discrepancy between perspectives, introducing an unsupervised non-learning approach to construct the intermediate states and aligning the two perspectives to the intermediate domain simultaneously. Extensive comparative experiments demonstrate that our method surpasses the state-of-the-art methods.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China National Science Fund for Distinguished Young Scholars under Grant 62025301, and the National Natural Science Foundation of China General Program under Grant 62473054, and the National Natural Science Foundation of China Basic Science Center Program under Grant 62088101.

References

Bansal, M.; Sawhney, H. S.; Cheng, H.; and Daniilidis, K. 2011. Geo-localization of street views with aerial image databases. In *Proceedings of the 19th ACM international conference on Multimedia*, 1125–1128.

Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33: 9912–9924.

Castaldo, F.; Zamir, A.; Angst, R.; Palmieri, F.; and Savarese, S. 2015. Semantic cross-view matching. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 9–17.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Dai, M.; Hu, J.; Zhuang, J.; and 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, Z.; Wang, G.; Yuan, W.; Zhu, S.; and Tan, P. 2022. Cluster contrast for unsupervised person re-identification. In *Proceedings of the Asian Conference on Computer Vision*, 1142–1160.

Deng, F.; Zhang, L.; Gao, F.; Qiu, H.; Gao, X.; and Chen, J. 2020. Long-range binocular vision target geolocation using handheld electronic devices in outdoor environment. *IEEE Transactions on Image Processing*, 29: 5531–5541.

Ding, L.; Zhou, J.; Meng, L.; and Long, Z. 2020. A practical cross-view image matching method between UAV and satellite for UAV-based geo-localization. *Remote Sensing*, 13(1): 47.

Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X.; et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, 226–231.

Gao, F.; Deng, F.; Li, L.; Zhang, L.; Zhu, J.; and Yu, C. 2021. MGG: Monocular global geolocation for outdoor long-range targets. *IEEE Transactions on Image Processing*, 30: 6349–6363.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hu, S.; Feng, M.; Nguyen, R. M.; and Lee, G. H. 2018. Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7258–7267.

Li, G.; Qian, M.; and Xia, G.-S. 2024. Unleashing Unlabeled Data: A Paradigm for Cross-View Geo-Localization. *arXiv preprint arXiv:2403.14198*.

Li, H.; Xu, C.; Yang, W.; Yu, H.; and Xia, G.-S. 2024. Learning Cross-view Visual Geo-localization without Ground Truth. *arXiv preprint arXiv:2403.12702*.

Lin, J.; Zheng, Z.; Zhong, Z.; Luo, Z.; Li, S.; Yang, Y.; and Sebe, N. 2022. Joint representation learning and keypoint detection for cross-view geo-localization. *IEEE Transactions on Image Processing*, 31: 3780–3792.

Lin, T.-Y.; Belongie, S.; and Hays, J. 2013. Cross-view image geolocalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 891–898.

Liu, L.; and Li, H. 2019. Lending orientation to neural networks for cross-view geo-localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5624–5633.

Reinhard, E.; Adhikhmin, M.; Gooch, B.; and Shirley, P. 2001. Color transfer between images. *IEEE Computer graphics and applications*, 21(5): 34–41.

Senlet, T.; and Elgammal, A. 2011. A framework for global vehicle localization using stereo images and satellite and road maps. In *2011 IEEE international conference on computer vision workshops (ICCV Workshops)*, 2034–2041. IEEE.

Shen, T.; Wei, Y.; Kang, L.; Wan, S.; and 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*, 1–1.

Shetty, A.; and Gao, G. X. 2019. Uav pose estimation using cross-view geolocalization with satellite imagery. In *2019 International Conference on Robotics and Automation (ICRA)*, 1827–1833. IEEE.

Shi, Y.; Liu, L.; Yu, X.; and Li, H. 2019. Spatial-aware feature aggregation for image based cross-view geo-localization. *Advances in Neural Information Processing Systems*, 32.

Shi, Y.; Yu, X.; Campbell, D.; and Li, H. 2020a. Where am i looking at? joint location and orientation estimation by cross-view matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4064–4072.

Shi, Y.; Yu, X.; Liu, L.; Zhang, T.; and Li, H. 2020b. Optimal feature transport for cross-view image geo-localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11990–11997.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sun, B.; Chen, C.; Zhu, Y.; and Jiang, J. 2019. Geocapsnet: Ground to aerial view image geo-localization using capsule network. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, 742–747. IEEE.

Sun, B.; Liu, G.; and Yuan, Y. 2023. F3-Net: Multiview Scene Matching for Drone-Based Geo-Localization. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–11.

Tian, X.; Shao, J.; Ouyang, D.; and 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.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, T.; Zheng, Z.; Sun, Y.; Chua, T.-S.; Yang, Y.; and Yan, C. 2022. Multiple-environment Self-adaptive Network for Aerial-view Geo-localization. *arXiv preprint arXiv:2204.08381*.

Wang, T.; Zheng, Z.; Yan, C.; Zhang, J.; Sun, Y.; Zheng, B.; and 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.

Workman, S.; Souvenir, R.; and Jacobs, N. 2015. Wide-area image geolocalization with aerial reference imagery. In *Proceedings of the IEEE International Conference on Computer Vision*, 3961–3969.

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3733–3742.

Yang, H.; Lu, X.; and Zhu, Y. 2021. Cross-view geo-localization with layer-to-layer transformer. *Advances in Neural Information Processing Systems*, 34: 29009–29020.

Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; and Hoi, S. C. 2021. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6): 2872–2893.

Zhang, J.; Lin, L.; and Liu, J. 2023. Hierarchical consistent contrastive learning for skeleton-based action recognition with growing augmentations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 3427–3435.

Zhang, L.; Deng, F.; Chen, J.; Bi, Y.; Phang, S. K.; Chen, X.; and Chen, B. M. 2018. Vision-based target three-dimensional geolocation using unmanned aerial vehicles. *IEEE Transactions on Industrial Electronics*, 65(10): 8052–8061.

Zhang, L.; Gao, F.; Deng, F.; Xi, L.; and Chen, J. 2022. Distributed Estimation of a Layered Architecture for Collaborative Air–Ground Target Geolocation in Outdoor Environments. *IEEE Transactions on Industrial Electronics*, 70(3): 2822–2832.

Zheng, H.; Zhao, J.; Zhu, J.; Ye, Z.; and Deng, F. 2024. Long-term urban air quality prediction with hierarchical attention loop network. *Sustainable Cities and Society*, 106010.

Zheng, Z.; Wei, Y.; and 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, J.; Cai, S.; Deng, F.; Ooi, B. C.; and Wu, J. 2024. Do LLMs Understand Visual Anomalies? Uncovering LLM's Capabilities in Zero-shot Anomaly Detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 48–57.

Zhu, J.; Cai, S.; Deng, F.; Ooi, B. C.; and Zhang, W. 2023a. METER: A Dynamic Concept Adaptation Framework for Online Anomaly Detection. *arXiv preprint arXiv:2312.16831*.

Zhu, R.; Yin, L.; Yang, M.; Wu, F.; Yang, Y.; and Hu, W. 2023b. 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, S.; Shah, M.; and Chen, C. 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.

Zhu, S.; Yang, T.; and Chen, C. 2021. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3640–3649.