Enhancing Cross-View Geo-Localization With Domain Alignment and Scene Consistency

Panwang Xia, *Graduate Student Member, IEEE*, Yi Wan, *Member, IEEE*, Zhi Zheng, *Member, IEEE*, Yongjun Zhang, *Member, IEEE*, and Jiwei Deng

***Abstract*— Cross-View Geo-Localization task is aimed at establishing correspondences between images captured from different perspectives within the same geographical region. The major challenge lies in the significant appearance variations of the same scene in different views. Current methods predominantly rely on learning a representation of the coarse-grained information from images and then evaluating the similarity, while the fine-grained features are usually not well-treated. In this paper, a novel method, named DAC (Domain Alignment and scene Consistency) is proposed, which leverages contrastive learning to acquire the global information of images and simultaneously employs a domain space alignment module to align the fine-grained features. The comprehensive utilization of multi-grained vision information guarantees better feature representations. Additionally, a cross-batch scene consistency strategy is proposed in the network to establish the global supervision of the positive samples based on scene correspondence, which improves the distinctiveness of the image representations. Advanced performance is shown by our method in drone-view target localization and drone navigation applications, outperforming state-of-the-art methods in comprehensive tests on the popular public datasets University-1652 and SUES-200. Our method also outperforms existing methods in cross-region localization, showing an average improvement of 5.6% in the R@1. Our codes and models are available at https://github.com/SummerpanKing/DAC.**

***Index Terms*— Cross-view, contrastive representation learning, geo-localization, image retrieval.**

I. Introduction

CROSS-VIEW geo-localization aims to determine the location of an image within a large-scale environment by establishing correspondences between query images and GPS-tagged satellite reference images. This task is applicable across diverse fields including autonomous driving, robot navigation, and object/event detection [1], [2], [3], [4]. Cross-view geo-localization (CVGL) was initially developed as a supplement to GPS-based positioning to handle the ground-to-satellite image localization task, particularly in urban environments with tall buildings interfering with Global Navigation Satellite System (GNSS) signals [5]. In recent years, with the widespread adoption of low-altitude unmanned aerial vehicle (UAV) technology, researchers have been captivated by CVGL based on images captured from near-ground perspectives by drones.

Drone-view target localization and drone navigation are two basic application scenarios that were specified with the introduction of the first dataset for near-ground-drone-captured images in the CVGL task [1]. Drone images are used as query images in drone-view target localization to find the most comparable satellite images, which locates the target building in the satellite view. However, in drone navigation, satellite images serve as query images to find the most pertinent drone images, making it easier for the drone to navigate back to the intended area using its flight history. With the advancements in deep learning, research on CVGL based on "drone-satellite" pairs has made significant progress. Despite representing images as global features (coarse-grained features) during retrieval, most existing approaches utilize the local features (fine-grained features) of images during network training. For instance, [6], [7] employ handcraft or trained block-wise weighting of images to establish spatial connections at the patch level between cross-view images. [8] enhances the fine-grained features of images through triplet attention and combines information using a multi-head classifier to obtain robust image representations. However, existing methods, while incorporating the fine-grained features of images in geo-localization tasks, neglect the problem of fine-grained feature misalignment between drone and satellite image features due to viewpoint differences. This oversight results in a decrease in the accuracy of image representation, thereby impacting the performance of CVGL.

Therefore, we propose a new multi-grained network for CVGL, named DAC (Domain Alignment and Scene Consistency) Network, which focuses on effectively leveraging both the coarse and fine-grained features of images. Inspired by [9], the DAC network employs a contrastive learning approach to acquire knowledge from the coarse-grained information

Manuscript received 18 March 2024; revised 11 July 2024; accepted 8 August 2024. Date of publication 14 August 2024; date of current version 23 December 2024. This work was supported in part by China Railway Group Laboratory Basic Research Project under Grant L2023G014, in part by the National Natural Science Foundation of China under Grant 42030102, in part by the Major Special Projects of Guizhou under Grant [2022]001, and in part by the National Engineering Research Center for Digital Construction and Evaluation Technology of Urban Rail Transit under Grant 2023[014]. This article was recommended by Associate Editor M. Zhang. *(Corresponding authors: Yongjun Zhang; Yi Wan.)*

Panwang Xia, Yi Wan, and Yongjun Zhang are with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: xiapanwang@whu.edu.cn; yi.wan@whu.edu.cn; zhangyj@whu.edu.cn).

Zhi Zheng is with the Department of Geography and Resource Management, The Chinese University of Hong Kong, Hong Kong, China (e-mail: zhizheng@cuhk.edu.hk).

Jiwei Deng is with China Railway Design Group Company Ltd., Tianjin 300308, China (e-mail: dengjiwei@crdc.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2024.3443510.

Digital Object Identifier 10.1109/TCSVT.2024.3443510

of images. Unlike previous methods, when utilizing the fine-grained features of images, our method predicts the implicit geographical patterns for each pixel in the image's fine-grained features to establish explicit constraints between drone-view and satellite-view, achieving cross-view spatial alignment. Simultaneously, our method establishes consistency between positive samples in different batches during contrastive learning, enabling the model to better learn information between drone images from different perspectives during training, thereby improving model performance and robustness.

In summary, our main contributions in this paper can be included as follows:

- A novel feature domain space alignment module is introduced to establish extra cross-view fine-grained constraints, empowering the backbone network with improved image representation capabilities for CVGL tasks.
- A global scene identity mapping strategy is proposed to explore the shared attributes among the samples across the batches, which enhances the aggregation of the same-region-captured images during the training phase.
- State-of-the-art performance in both drone-view target localization and drone navigation tasks has been achieved on two public datasets, University-1652 and SUES-200. Moreover, superior performance in cross-region transferability has been achieved, with an average improvement of 5.6% in the R@1 compared to existing methods.

The remaining sections of this paper are structured as follows: Section II provides a brief overview of related works, Section III delves into the details of our proposed DAC method, Section IV presents a comprehensive summary of experimental results, and Section V concludes the paper.

II. Related Works

A. Cross-View Geo-Localization

Due to its extensive application potential, the cross-view geo-localization task has long captured the attention of many scholars. Cross-view geo-localization primarily tackles two issues: geo-localization based on image retrieval in large-scale areas and precise geo-localization based on pose estimation in small-scale areas [10], [11], [12]. Our work is dedicated to addressing the former issue.

1) Datasets for Cross-View Geo-Localization: In earlier works [5], [13], scholars investigated the localization problem between ground-view and aerial-view images, organizing datasets into image pairs. Subsequently, scholars continued with this setup and created two large-scale geo-localization datasets, CVUSA [14] and CVACT [15], where ground-view panoramic images served as queries, and reference images were downward satellite images with GPS tags. In practical scenarios, query images do not precisely match the center of a reference image. Zhu et al. [16] introduced the Vigor geo-localization dataset, allowing the study of this problem under a more realistic and practical setting, serving as a testbed to bridge the gap between current research and practical applications. Later, Zheng et al. [1] proposed the University-1652 dataset to study cross-view geo-localization, introducing a new multi-view, multi-source dataset, including synthetic drone, satellite, and ground camera images. This provided feasibility validation for drone platforms in the field of geo-localization and prompted the introduction of two new tasks: drone-view target localization and drone navigation. Recently, Zhu et al. [17] designed the SUES-200 dataset to assess model adaptability in complex and changing scenarios, focusing on multiple heights, scenes, and continuous scenes. Wang et al. [18] expanded the University-1652 dataset to encompass multiple environmental conditions, aiming to evaluate the model's robustness.

2) Methods for Cross-View Geo-Localization: In the cross-view geo-localization task, extracting viewpoint-invariant image features is a critical challenge. Significant viewpoint differences between query and reference images lead to unavoidable feature domain disparities. In early works, researchers mitigated viewpoint differences between ground and aerial images by transforming ground images acquired from stereo cameras into bird's-eye-view (BEV) perspective images [19]. With the development of deep learning, researchers started utilizing convolutional neural networks (CNNs) as feature extraction tools. Workman and Jacobs [20] initially used a CNN model pre-trained on the ImageNet and Caffe datasets to extract deep image features, achieving cross-view geo-localization—furthermore, Workman et al. [14] fine-tuned neural networks on cross-view geo-localization datasets, enabling networks to learn shared semantic features across different view sources for better performance. Inspired by the success of neural networks in image classification tasks, Tian et al. [5] performed target geo-localization on multiple buildings in query images, considering the geographical topological relationships between buildings for image localization. Considering the geometric relationships between cross-view images, Liu and Li [15] introduced orientation encoding between ground query images and satellite reference images, achieving better geo-localization results. Additionally, Shi et al. [21] first proposed polar transform for satellite images to mitigate viewpoint differences, a technique later adopted by many works. However, polar transform introduces image distortion and loss of sky information, posing challenges to geo-localization. To address this, Shi et al. [22] used a GAN model to repair image distortions resulting from polar transform. In recent years, with the enhanced representational capabilities of neural networks, and considering that geometric properties of aerial images are disrupted after polar transform, Zhu et al. [23] used Vision Transformer as an image feature extractor, achieving excellent geo-localization results without geometric preprocessing. Most recently, Deuser et al. [9] used the powerful ConvNeXt network as a backbone, applying contrastive learning for cross-view geo-localization representation learning, achieving the current state-of-the-art performance.

In the context of drone-satellite cross-view geo-localization, as there is no fixed geometric relationship between drones and satellites, methods such as polar transform cannot be directly applied to eliminate viewpoint differences. Wang et al. [6] proposed that different positions in the drone view have varying impacts on localization, effectively improving the method's performance through manually designed block-wise weighting of drone images. Building upon this, Dai et al. [7] replaced the manually designed block-wise weighting in [6] with an automatic block-wise weighting approach based on a Transformer model, enhancing model robustness. Wang et al. [24] identified noise interference in image features used for localization and improved representational capabilities by introducing a dynamic weighted decorrelation regularization method. Shen et al. [8] utilized a ConvNeXt-base network as the main feature extraction model for drone-satellite geo-localization, achieving improved geo-localization through a three-channel classification approach.

Many previous efforts have attempted to leverage fine-grained features of images as auxiliary information to optimize geo-localization at the level of image space hierarchy [25]. However, current methods have not effectively addressed the spatial misalignment issue between query and reference images due to viewpoint differences. In this paper, our proposed DAC implements a domain space alignment module to achieve the alignment of the fine-grained image features at the pixel level.

B. Contrastive Representation Learning

The quest for enhancing the image encoding capability of neural networks has been a focal point in representation learning research, with a recent surge in the development of contrastive learning methods significantly boosting the representational power of neural networks. In earlier studies [26], researchers observed that images from closely related categories in image classification tasks share similar visual appearances (e.g., images of leopard and jaguar). This observation suggests that networks tend to focus more on visual information than semantic labels during image classification. Building upon this assumption, the authors adopted an extreme approach, treating each image as a separate class. Contrastive representation learning relies on the assistance of a proxy task during actual training. The aforementioned works are based on discriminative models, while [27] introduces a generative model-based contrastive representation learning method. This approach accepts multimodal data such as text, audio, and images as input and utilizes auto-regression to generate future predictions using past input information. Positive samples correspond to future inputs, while negative samples are arbitrarily chosen inputs. In addition to using the augmented results of the data itself as positive samples, the work by Tian et al. [28] proposes using scenes as the defining unit of samples. They argue that images from different modalities within the same scene should be considered positive samples due to the strong commonality between them.

Inspired by the aforementioned works, contrastive representation learning has also made significant progress recently. He et al. introduced MoCo [29], incorporating a momentum encoder module that surpasses the training effectiveness of supervised methods in unsupervised contrastive learning. Chen et al. proposed SimCLR [30], introducing a projector composed of a simple MLP to eliminate domain differences between the features of the two contrastive learning branches, resulting in significant performance improvement. Subsequently, MoCo v2 [31] and SimCLR v2 [32] were introduced, optimizing data augmentation, learning rate schedules, and momentum encoders to achieve better results.

Contrastive representation learning based on discriminative models requires simultaneous positive and negative sample constraints to ensure that the network model does not collapse during training. However, Grill et al. [33] set the contrastive learning proxy task as a generative model and introduced an MLP projector to establish mappings between branches. This configuration allows the network to break free from dependence on negative samples, enhancing training efficiency and stability. In subsequent works, researchers in [34] also adopted this structure, and [35], [36], [37] incorporated Vision Transformer into contrastive representation learning tasks.

In CVGL, representation methods based on contrastive learning prove instrumental in enabling networks to acquire excellent representational capabilities. However, within the existing methods, some noteworthy characteristics should be noticed: 1) In the training of contrastive learning networks, the presence of as many negative samples as possible within the same batch is crucial for the network to acquire discriminative representational capabilities [29]. 2) Building on the insights from [28], researchers propose treating shared information between different perspectives (or modalities) within the same scene as positive samples, advocating for its full extraction and utilization. 3) In the context of the CVGL task, multiple images from various perspectives are commonly available for the same geographical scene.

Guided by the aforementioned characteristics, the contrastive representation learning network designed for CVGL guarantees that any two images within the same batch do not originate from the same scene during training, which results in information dissipation. To address the problem, in this paper, we introduced a cross-batch scene consistency strategy, to build communication between images from the same region by a global geographic tag.

III. Proposed Method

In this section, we introduce the proposed DAC method, as depicted in Fig. 1. Our method comprises three components. First, a backbone network is used to extract global information from the input images (coarse-grained feature), employing a contrastive learning approach with symmetric InfoNCE as the loss function. Second, we focus on the cross-view fine-grained feature constraints, where an alignment module is introduced to align the query and the reference images in a shared feature domain. Thus, the domain gaps arising from viewpoint differences are mitigated. Third, a cross-batch scene consistency module is proposed. Beyond discerning differences between negative samples, this strategy aims to exploit similarities among the same-region samples across different batches. Notably, DAC is designed based on multi-metric optimization, and the proposed modules are not used during inference, resulting in no additional computational cost.

Problem formulation: Given a geo-localization dataset with geographical region labels denoted as i, and $i \in R$, where R is the number of geographical regions in the reference database.

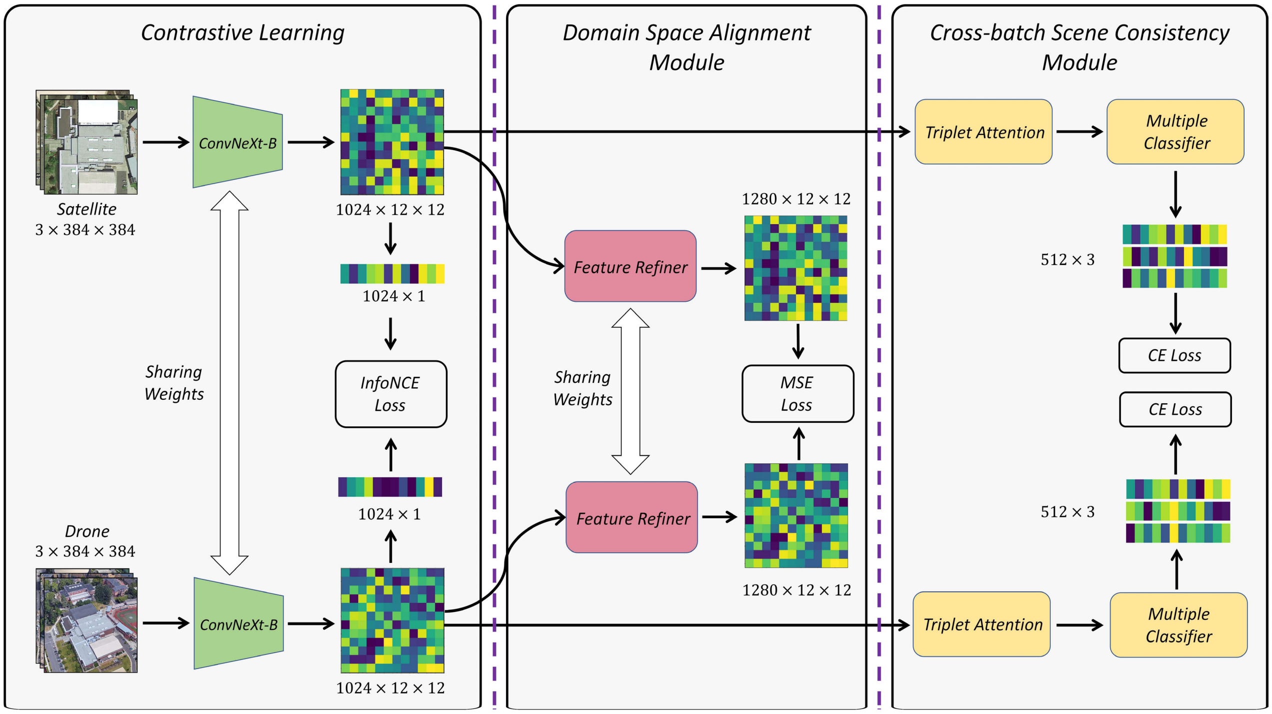


Fig. 1. Overview of our proposed DAC method. Images from different platforms are first fed into the ConvNeXt-B backbone to generate deep features. The Contrastive Learning part uses a symmetric InfoNCE loss to learn discriminative features in coarse-grained level. The Domain Space Alignment module focuses on better utilizing the fine-grained features by establishing cross-view constraints. The Cross-batch Scene Consistency module leverages the consistency of images from the same scene to enhance the representation ability (see text for details).

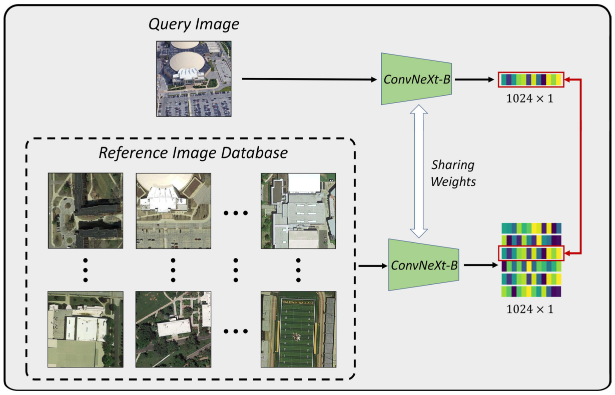


Fig. 2. The scheme of image retrieval-based cross-view geo-localization.

Let the query images be X_j^i, $j \in N$, where N is the number of images within one region. The reference images are denoted as Y^i. Our goal is to design and train a network that serves as a mapping function to represent images from different platforms X and Y into a shared feature space. In this feature space, images from the same geographical region should be closer, while images from different geographical regions should be farther apart.

The process of the CVGL task is illustrated in Fig. 2, where query images and reference images are encoded into embeddings by the backbone neural network. Localization is then determined based on the distance between the embeddings of the query and reference images.

A. Contrastive Learning Using ConNeXt Backbone

Contrastive learning methods assist neural network models in acquiring knowledge by comparing the similarity among positive samples and the dissimilarity between negative samples. The objective is to learn distinctive representational properties for each class of samples. In the context of cross-view geo-localization tasks, the feature domain gap arising from significant viewpoint differences between query and reference images poses a critical challenge to successful localization. Inspired by previous works such as [9] and [38], we adopt a contrastive learning approach to train the geo-localization network. A ConvNeXt model pre-trained on ImageNet is used as the backbone for the geo-localization network. To learn better cross-view patterns within the geo-localization dataset, a contrastive learning method based on a discriminative model, training with positive and negative samples, is adopted. The symmetric InfoNCE loss is employed as the training loss to supervise the network, as indicated in (1):

$$\mathcal{L}_{InfoNCE}(q, R) = -log\frac{\exp\left(q \cdot \frac{r_+}{\tau}\right)}{\sum_{i=0}^{R}\exp\left(q \cdot \frac{r_i}{\tau}\right)} \tag{1}$$

The InfoNCE loss serves as a commonly used contrastive loss function, offering better robustness compared to traditional cross-entropy loss functions by considering negative sample information. Here, q represents the query image, and R represents a set of reference images, where the positive sample corresponding uniquely to the query image q is denoted as r_+. In a training batch, if the value of $q{\cdot}r_+$ between positive samples is higher and between negative samples is lower, $\mathcal{L}(q, R)_{InfoNCE}$ approaches 0; otherwise, $\mathcal{L}(q, R)_{InfoNCE}$ grows exponentially. The hyperparameter τ, also known as the temperature parameter, can be a learned or fixed value.

As indicated by (1), the InfoNCE loss guides the model to enhance the similarity of the positive sample feature

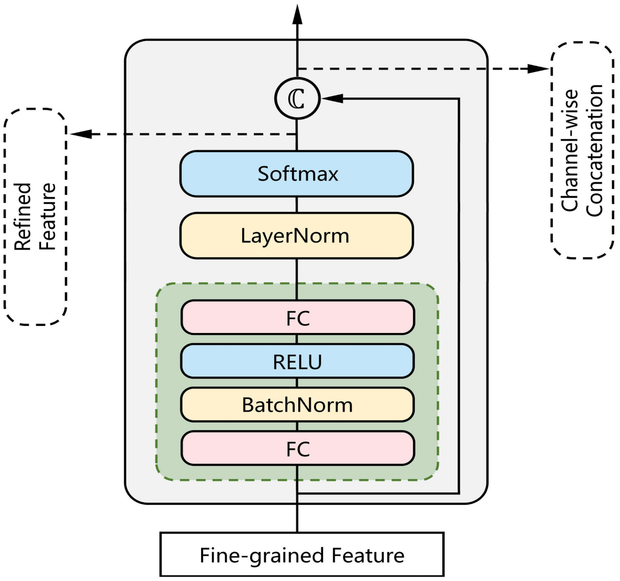


Fig. 3. The process of the Feature Refiner in the DSA module. The fine-grained features of images are sent to a MLP block followed by normalization and activation, which can extract sub-features specially for cross-view alignment. Then a channel-wise concatenation is utilized to get the combined fine-grained features.

representations and enhance the dissimilarity between pairs of negative samples. This mechanism forces the model to overcome the interference of the cross-view variations. Additionally, the DAC network employs symmetric InfoNCE loss in computations, which simultaneously calculates losses for query and reference features. This bidirectional loss constraint accelerates model convergence and facilitates the learning of the enhanced image feature representations.

B. Domain Space Alignment Module

In the cross-view geo-localization task, relying solely on the global (coarse-grained) information of images is insufficient to tackle the challenges of cross-view variations. Existing cross-view geo-localization methods also face challenges in effectively leveraging both the coarse-grained and fine-grained information within images (See Section II-A). Therefore, inspired by LEWEL [39], we introduce a novel domain space alignment (DSA) module, as depicted in Fig. 1.

The DSA module, as part of the multi-metric optimization, establishes explicit constraints between cross-view fine-grained features during training. These constraints help the backbone network overcome the domain gap caused by cross-view differences and learn more discriminative features. The DSA module consists of two components: the Feature Refiner and the Fine-grained channel-wise concatenation.

1) Feature Refiner: ConvNeXt is a fully convolutional backbone neural network that contains the spatial attributes of the original image in its output fine-grained features. The $F_q, F_r \in \mathbb{R}^{h \times w \times d_F}$ are denoted to be fine-grained features generated by backbone network from query and reference images, where $d_F = 1024$. The F features contain the whole information from the images, but not all of them are required for domain space alignment. To better utilize the fine-grained information for cross-view alignment, we employ a simple MLP as the Feature Refiner. The Feature Refiner is designed to extract features specially for alignment from the origin fine-grained features, and we named them refined features f_q, f_r, with dimensions $(h \times w \times d_f)$, where $d_f = 256$. Compared to the backbone network, the additional MLP Refiner has limited parameters, which leverages fine-grained spatial information while ensure training stability at the same time. After extraction by the MLP Feature Refiner, each position in the fine-grained feature is assigned a 256-dimensional feature, serving as an implicit semantic representation for alignment.

2) Fine-Grained Channel-Wise Concatenation: After Feature Refiner, we concatenate F and f along the channel dimension. The reasons that we choose concatenation rather than multiplication can be summarized as follows: 1) Unlike LEWEL, which emphasizes "where is more important," we prioritize cross-view correspondence in spatial dimensions. The fine-grained alignment matrices generated by the MLP should be viewed as refined features, especially for alignment purposes (similar to pixel-wise semantic segmentation of the feature map), rather than as weights. 2) We aim to ensure that the MLP refiner does not surpass the capabilities of the backbone network as the MLP is not part of the model during inference. Thus, it is natural for us to maintain the independence of the original representations by using concatenation.

The concatenation results in the combined fine-grained features, $FC_q, FC_r \in \mathbb{R}^{h \times w \times d_{FC}}$, where $d_{FC} = 1280$. The loss function measuring the similarity between features of cross-view images is formulated as Mean Squared Error (MSE):

$$\mathcal{L}_{MSE}(FC_q, FC_r) = 1 - \frac{1}{M}\sum_{i=1}^{M} FC_q \cdot FC_r, \qquad (2)$$

where M represents the number of images, FC_q and FC_r denote the combined fine-grained features obtained from drone and satellite images respectively. The combined fine-grained features FC_q and FC_r are then used to calculate the loss after normalization. During training, both F and f can achieve growth in the alignment capabilities of the backbone, and the growth of f can also help with the growth of F's capabilities.

C. Cross-Batch Scene Consistency Module

The key to addressing the challenge of cross-view geo-localization lies in extracting distinctive features from images captured from different perspectives. However, as illustrated in the left box of Fig. 4, InfoNCE loss can only capture the positive/negative sample relationships within the current batch, while ignoring the relationship between images belonging to the same scene in the different batches (See Section II-B).

To bridge the communication of positive samples (drone-view images in the same scene) in different batches, as illustrated in the right box of Fig. 4, we introduce the cross-batch scene consistency (CSC) strategy, using an additional global cross-entropy loss during training in our proposed DAC method.

Features extracted by the backbone network are fed into the cross-batch scene consistency component. Inspired by the

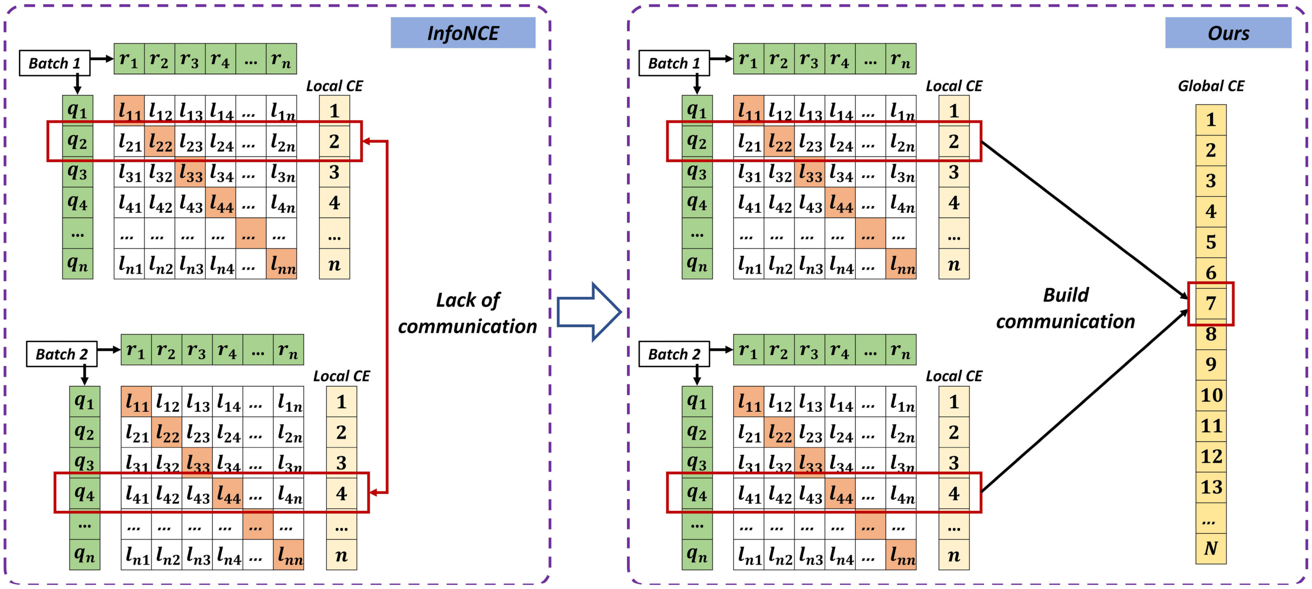

InfoNCE
Batch 1
r_1 r_2 r_3 r_4 … r_n
q_1 q_2 q_3 q_4 … q_n
l_{11} l_{12} l_{13} l_{14} … l_{1n}
l_{21} l_{22} l_{23} l_{24} … l_{2n}
l_{31} l_{32} l_{33} l_{34} … l_{3n}
l_{41} l_{42} l_{43} l_{44} … l_{4n}
l_{n1} l_{n2} l_{n3} l_{n4} … l_{nn}
Local CE
1 2 3 4 … n
Batch 2
Lack of communication
Ours
Global CE
1 2 3 4 5 6 7 8 9 10 11 12 13 … N
Build communication

Fig. 4. The process of the cross-batch scene consistency. Compared to existing contrastive loss function (e.g., InfoNCE, shown in the left block), our strategy (shown in the right block) establishes the communication between q_2 and q_4 in different batch but belonging to the same region, aiding the network to learn better representation ability.

existing works, we apply a triplet attention [40] module to enhance the fine-grained features F_{drone} and $F_{satellite}$ before extracting global features. Then, a multi-head classifier [8] module is utilized for better encoding, and a cross-entropy loss is employed as the global supervision.

$$\mathcal{L}_{CE}\left(F, label\right) = -\log\left(\frac{exp\left(F\left[label\right]\right)}{\sum_j exp\left(F\left[j\right]\right)}\right), \tag{3}$$

where F represents the feature map of satellite or drone images, while $label$ denotes the geographical region label corresponding to the image, the set of j encompasses all scene labels.

With this strategy, the DAC method establishes communication between the different-view-but-same-region images across different batches, which forces the network to focus on both learning differences among negative samples and recognizing similar patterns among positive samples. Compared to the original InfoNCE loss, the shared information among different perspectives is fully exploited, which effectively aids the model in better representation.

D. Loss Function

In our training process, the three components of the DAC method provide supervisory constraints for model training:

$$\mathcal{L}oss = \mathcal{L}_{InfoNCE} + \lambda_1 \mathcal{L}_{MSE} + \lambda_2 \mathcal{L}_{CE}, \tag{4}$$

where λ_1 and λ_2 are adjustable hyperparameters. $\mathcal{L}_{InfoNCE}$ forces the network to learn distinctive feature representations of different geographical scenes during training. $\mathcal{L}_{MSE}$ forces the network to better focus on the correspondences between images from different perspectives, which helps the network to learn effective features. $\mathcal{L}_{CE}$ complements the shortcomings of $\mathcal{L}_{InfoNCE}$, compelling the network to attain better feature representations among positive samples of drone-view images.

IV. Experiments

In Section IV-A, we introduce the two mainstream public cross-view geo-localization datasets used in our experiments, along with the performance evaluation metrics employed for our method. In Section IV-B, we provide detailed information on the parameter configurations and experimental setup. Section IV-C comprehensively compares our method with existing state-of-the-art methods. Section IV-D focuses on evaluating our method with existing state-of-the-art methods in cross-region transferability. Finally, in Sections IV-E and IV-F, we present the results of the conducted ablation experiments and visualizations, respectively.

A. Datasets and Evaluation Protocols

Our method is designed to address the cross-view geo-localization challenge between drone-view and satellite-view images. We trained and tested our model on three mainstream datasets, University-1652 [1], SUES-200 [17] and Multi-weather University-1652 [18], which are widely utilized in recent state-of-the-art works [6], [7], [8] on drone-satellite geo-localization.

University-1652 is a cross-view geo-localization dataset released in 2020, encompassing imagery from three distinct elevation platforms: simulated drones, satellites, and ground-level perspectives capturing natural scenes of 1652 university buildings worldwide. Notably, University-1652 stands out as the pioneering cross-view geolocation dataset based on drone-view images, establishing itself as a prominent benchmark in current research endeavors. The dataset comprises 701 buildings from 33 universities in the training set and 951 buildings from 39 universities in the test set, with no overlap between buildings in the two sets.

SUES-200 is a relatively recent cross-view dataset, released in 2023. This dataset comprises images from two perspectives:

TABLE I
COMPARISON WITH STATE-OF-THE-ART RESULTS ON UNIVERSITY-1652. THE BEST RESULTS ARE IN BOLD

Method	Drone→Satellite		Satellite→Drone	
	R@1	AP	R@1	AP
LPN TCSVT'22 [6]	75.93	79.14	86.45	74.79
FSRA TCSVT'22 [7]	82.25	84.82	87.87	81.53
MSBA RemoteSensing'21 [41]	82.33	84.78	90.58	81.61
Swin-B+DWDR arxiv'22 [24]	86.41	88.41	91.30	86.02
MBF Sensors'23 [42]	89.05	90.61	93.15	88.17
MCCG TCSVT'23 [8]	89.64	91.32	94.30	89.39
Sample4Geo ICCV'23 [9]	92.65	93.81	95.14	91.39
DAC (Ours)	**94.67**	**95.50**	**96.43**	**93.79**

drone-view and satellite-view. What sets SUES-200 apart from previous datasets is its inclusion of tilt images from the drone-view perspective with labeled flight heights (150 meters, 200 meters, 250 meters, and 300 meters), bringing it closer to real-world scenarios. Furthermore, this dataset captures a diverse range of scene types, extending beyond campus buildings to encompass parks, schools, lakes, and public structures.

Multi-weather Univeraity-1652 is a newly proposed dataset based on the University-1652. Ten well-designed augmentation settings are utilized to simulate the multiple environments including fog, rain, snow and other adverse weathers.

1) Evaluation Protocols: In our experiments, we employ two widely-used performance metrics, Recall@K (R@K) and Average Precision (AP), to evaluate the effectiveness of our model. These metrics have gained widespread usage in existing literature. R@K signifies the proportion of correctly matched images within the top K ranked results of the localization output. A higher R@K score is indicative of superior performance by the network model. Additionally, we calculate the area under the Precision-Recall curve, known as Average Precision (AP), to quantify the precision and recall rates of the retrieval performance. These metrics collectively provide a comprehensive assessment of the model's accuracy and efficiency in the context of our experiments.

B. Implementation Details

We employed the ConvNeXt-Base model pretrained on the ImageNet-22k dataset as our backbone network. The network parameters in the domain space alignment and the cross-batch scene consistency were initialized using the Kaiming initialization method. During model training, images were resized to 384 × 384 [8], [9] and subjected to a series of image augmentation processes, including color space adjustments, blur, sharpening, random cropping, and random occlusion. Batchsize was set to 24 (24 drone images and 24 satellite images each, equivalent to 48 images). The AdamW optimizer with an initial learning rate of 0.001 was utilized, and a cosine scheduler was employed with the first 10% of training period as warm-up. Hyperparameters of the loss function in DAC are set as $\lambda_1 = 0.6$, $\lambda_2 = 0.1$. Our network model was implemented using the PyTorch platform, and experiments were conducted on a desktop computer running Ubuntu 22.04. The desktop computer is equipped with an NVIDIA GeForce RTX 4090 GPU with 24GB of memory.

C. Comparison With the State-of-the-Art Methods

1) Results on University-1652: As shown in the Table I, our proposed DAC method achieves an R@1 of 94.67% and an AP of 95.50% in Drone→Satellite task, and an R@1 of 96.43% and an AP of 93.79% in Satellite→Drone task. With the help of the proposed modules, the DAC method shows better representation ability compared to the state-of-the-art methods.

2) Results on SUES-200: We also tested our proposed DAC method on the SUES-200 dataset. As shown in the Table II, the results demonstrate that the DAC method achieves state-of-the-art performance and exhibits robustness when faced with variations in drone platform image altitudes.

3) Results on Multi-weather University-1652: To further evaluate the robustness of the DAC method under multi-weather conditions, we have also conducted experiments on the Multi-weather University-1652 dataset. As shown in Table III, the results indicate that the proposed DAC method can effectively operate in real-world scenarios, including adverse weather conditions.

D. Comparison With the State-of-the-Art Methods in Cross-Region Transferability

The transferability of cross-view geo-localization methods is an intriguing and practically significant metric. Transferability refers to the ability of a model trained in one region to perform geo-localization in another unseen region with different styles. To assess the transferability of our proposed method, we designed experiments using the training set of the University-1652 dataset as the training data and the test set of the SUES-200 dataset as the testing data (since the multi-height organization of the SUES-200 dataset results in a limited number of scenes, training on SUES-200 was not conducted to test transferability). In the experiment, we selected the current state-of-the-art methods for "drone-satellite" cross-view geo-localization, MCCG and Sample4Geo, for comparison. All methods employed the same settings (using ConvNeXt-base as the backbone, image input size of 384, and training batch size equivalent to 48). The results in the Table IV show that, while not achieving the

TABLE II
COMPARISON WITH STATE-OF-THE-ART RESULTS ON SUES-200. THE BEST RESULTS ARE IN BOLD

Drone→Satellite								
Method	150m		200m		250m		300m	
	R@1	AP	R@1	AP	R@1	AP	R@1	AP
SUES-200 Baseline TCSVT'23 [17]	55.65	61.92	66.78	71.55	72.00	76.43	74.05	78.26
LPN TCSVT'22 [6]	61.58	67.23	70.85	75.96	80.38	83.80	81.47	84.53
FSRA TCSVT'22 [7]	68.25	73.45	83.00	85.99	90.68	92.27	91.95	91.95
MCCG TCSVT'23 [8]	82.22	85.47	89.38	91.41	93.82	95.04	95.07	96.20
Sample4Geo ICCV'23 [9]	92.60	94.00	97.38	97.81	**98.28**	**98.64**	**99.18**	**99.36**
DAC (Ours)	**96.80**	**97.54**	**97.48**	**97.97**	98.20	98.62	97.58	98.14
Satellite→Drone								
Method	150m		200m		250m		300m	
	R@1	AP	R@1	AP	R@1	AP	R@1	AP
SUES-200 Baseline TCSVT'23	75.00	55.46	85.00	66.05	86.25	69.94	88.75	74.46
LPN TCSVT'22	83.75	66.78	88.75	75.01	92.50	81.34	92.50	85.72
FSRA TCSVT'22	83.75	76.67	90.00	85.34	93.75	90.17	95.00	92.03
MCCG TCSVT'23	93.75	89.72	93.75	92.21	96.25	96.14	98.75	96.64
Sample4Geo ICCV'23	**97.50**	93.63	**98.75**	**96.70**	**98.75**	**98.28**	**98.75**	**98.05**
DAC (Ours)	**97.50**	**94.06**	**98.75**	96.66	**98.75**	98.09	**98.75**	97.87

TABLE III
COMPARISON WITH STATE-OF-THE-ART RESULTS ON MULTI-WEATHER UNIVERSITY-1652. THE BEST RESULTS ARE IN BOLD

Method	Normal	Fog	Rain	Snow	Fog+Rain	Fog+Snow	Rain+Snow	Dark	Over -exposure	Wind
	R@1/AP	R@1/AP	R@1/AP	R@1/AP	R@1/AP	R@1/AP	R@1/AP	R@1/AP	R@1/AP	R@1/AP
Drone→Satellite										
MuSeNet[18]	74.48/77.83	69.47/73.24	70.55/74.14	65.72/69.70	65.59/69.64	54.69/59.24	66.64/70.55	53.85/58.49	61.05/65.51	69.45/73.22
Sample4Geo	90.55/92.18	89.72/91.48	85.89/88.11	86.64/88.81	85.88/88.16	84.64/87.11	85.65/87.93	87.15/89.32	76.72/80.18	83.39/89.51
DAC	**92.81/94.06**	**92.55/93.84**	**90.06/91.70**	**90.04/91.69**	**89.80/91.45**	**89.29/91.02**	**89.91/91.52**	**91.00/92.54**	**82.41/85.18**	**90.98/92.48**
Satellite→Drone										
MuSeNet	88.02/75.10	87.87/69.85	87.73/71.12	83.74/66.52	85.02/67.78	80.88/54.26	84.88/67.75	80.74/53.01	81.60/62.09	86.31/70.03
Sample4Geo	95.86/89.86	95.72/88.95	94.44/85.71	95.01/86.73	93.44/85.27	93.72/84.78	93.15/85.50	96.01/87.06	89.87/74.52	95.29/87.06
DAC	**97.43/92.90**	**97.00/92.44**	**95.72/90.00**	**96.01/90.48**	**95.29/89.52**	**95.86/89.40**	**95.29/89.68**	**96.86/90.84**	**94.15/81.80**	**95.86/90.97**

geo-localization performance after training in the same scene, our proposed DAC method still achieved a remarkably high localization success rate.

Furthermore, compared to the current state-of-the-art methods, DAC outperforms by a large margin in geo-localization performance. Particularly when there is a significant viewpoint difference between query and reference images (height at 150 meters), DAC, compared to Sample4Geo, achieves a 6.60% increase in R@1 and 5.63% in AP for the drone-view target localization task (Drone→Satellite) and a 3.75% increase in R@1 and 6.04% in AP for the drone navigation task (Satellite→Drone). These results demonstrate that our method, based on contrastive learning using multi-grained information, contributes to the performance improvement of cross-view geo-localization models.

E. Ablation Studies

In the ablation studies, we investigated the effects of the cross-batch scene consistency and domain space alignment proposed in this paper. These strategies are designed to assist the geo-localization network in establishing correlations between images during training and overcoming challenges posed by cross-view variations. To validate the effectiveness of the proposed modules, we conducted experiments both with and without the inclusion of these strategies under identical conditions.

1) Effect of the Cross-Batch Scene Consistency Module: We abbreviate the cross-batch scene consistency as CSC. As shown in Table V, incorporating CSC in the network enhances model performance. Compared to the model without CSC, our method achieves a performance improvement of 1.93% in R@1 and 1.52% in AP for the drone-view target localization task (Drone→Satellite). For the drone navigation task (Satellite→Drone), there is a performance gain of 0.15% in R@1 and 1.32% in AP. CSC is designed to extract shared information among multi-perspective drone images within the same geographical scene. This is evident from the significant performance improvement in R@1 and AP for the drone-view

TABLE IV
COMPARISON WITH STATE-OF-THE-ART RESULTS IN CROSS-REGION TRANSFERABILITY. THE BEST RESULTS ARE IN BOLD

Drone→Satellite								
Method	150m		200m		250m		300m	
	R@1	AP	R@1	AP	R@1	AP	R@1	AP
MCCG [8]	57.62	62.80	66.83	71.60	74.25	78.35	82.55	85.27
Sample4Geo [9]	70.05	74.93	80.68	83.90	87.35	89.72	90.03	91.91
DAC (Ours)	**76.65**	**80.56**	**86.45**	**89.00**	**92.95**	**94.18**	**94.53**	**95.45**
Satellite→Drone								
Method	150m		200m		250m		300m	
	R@1	AP	R@1	AP	R@1	AP	R@1	AP
MCCG	61.25	53.51	82.50	67.06	81.25	74.99	87.50	80.20
Sample4Geo	83.75	73.83	91.25	83.42	93.75	89.07	93.75	90.66
DAC (Ours)	**87.50**	**79.87**	**96.25**	**88.98**	**95.00**	**92.81**	**96.25**	**94.00**

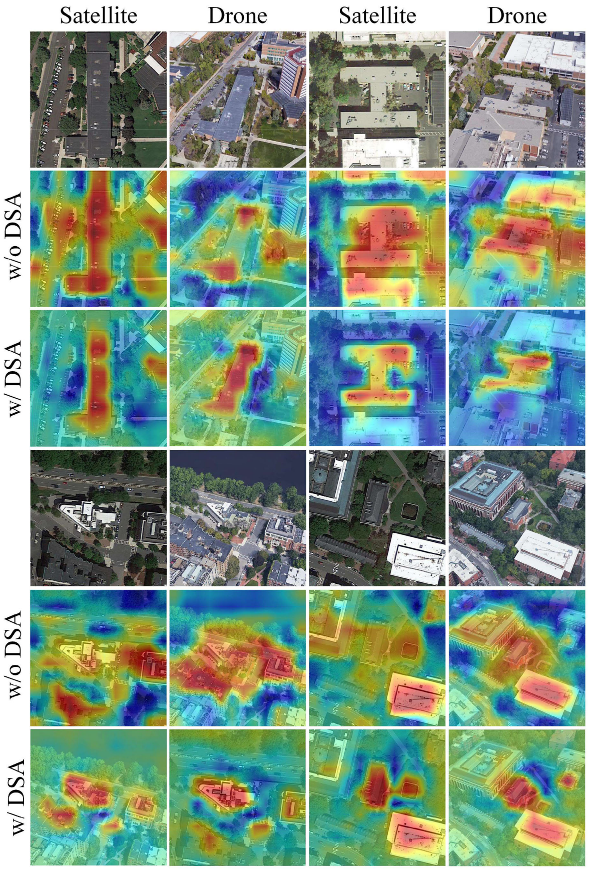
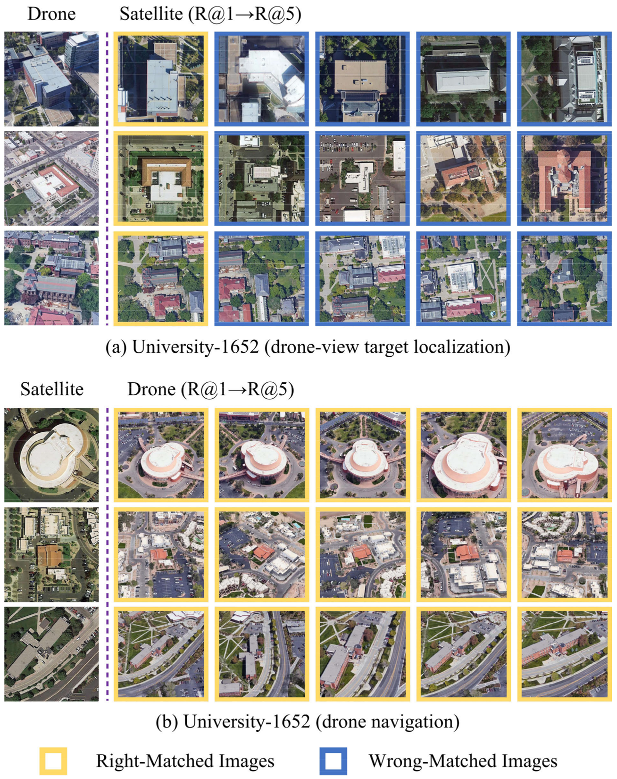


Fig. 5. The heatmaps of the fine-grained features with or without the DSA module. 4 pairs of drone-satellite images in the University-1652 are chosen to illustrate the domain gap erasing ability. Heatmaps on cross-view images represent the correlation between them through similar colors.



(a) University-1652 (drone-view target localization)

(b) University-1652 (drone navigation)

Fig. 6. Cross-view geo-localization results. (a) Top-5 results of drone-view target localization. (b) Top-5 results of drone navigation. The yellow box represents a right-matched image, and the blue box represents a wrong-matched image.

target localization task. Even in the drone navigation task, where the baseline performance is already high (matching a satellite-view image to any of dozens of drone-view images is considered successful), the network with the CSC module achieves a notable improvement, emphasizing the effectiveness of CSC.

2) Effect of the Domain Space Alignment Module: We abbreviate domain space alignment as DSA. DSA is designed to address the cross-view challenges between drone and satellite images by achieving feature domain alignment through acquiring implicit semantic representations of fine-grained image features. As indicated in the results in Table V, incorporating DSA in the geo-localization network, compared to not using this module, results in a 2.42% increase in R@1 and a 1.98% increase in AP for the drone-view target localization task (Drone→Satellite). In the drone navigation task (Satellite→Drone), it achieves an improvement of 0.72% in R@1 and 1.99% in AP. The significant performance improvement strongly demonstrates the effectiveness of the proposed DSA.

TABLE V
THE RESULTS OF ABLATION STUDIES OF PROPOSED METHOD ON UNIVERSITY-1652

Method	Drone→Satellite		Satellite→Drone	
	R@1	AP	R@1	AP
w/o CSC, DSA	91.69	93.10	95.29	91.16
w/ CSC	93.62 ↑ 1.93	94.62 ↑ 1.52	95.44 ↑ 0.15	92.48 ↑ 1.32
w/ DSA	94.11 ↑ 2.42	95.08 ↑ 1.98	96.01 ↑ 0.72	93.15 ↑ 1.99
DAC (Ours)	**94.67** ↑ 2.98	**95.50** ↑ 2.40	**96.43** ↑ 1.14	**93.79** ↑ 2.63

Additionally, we conducted a qualitative analysis of the DSA module. As depicted in Fig. 5, the heatmaps are derived from fine-grained features generated by the backbone model trained with and without the DSA module. The results show that the DSA module helps the backbone focus on objects more precisely under cross-view conditions, thereby improving cross-view alignment.

F. Visualization of Qualitative Results

As an additional qualitative evaluation, we present the visualization of retrieval results for different tasks on the University-1652 dataset. In the drone-view target localization task (depicted in Fig. 6(a)) and drone navigation task (depicted in Fig. 6(b)), we randomly selected three scenes for each task and displayed the top five retrieval results based on the model's computations. Correctly matched results between query and reference images are highlighted in yellow boxes, while incorrect results are marked with blue boxes. The results indicate that the DAC method is applicable to cross-view geo-localization tasks.

V. CONCLUSION

In this paper, we investigate the cross-view geo-localization task between drone-view images and satellite-view images. We proposed a cross-view geo-localization method, named DAC, which utilized a domain space alignment module to overcome the cross-view differences between drone and satellite images and employs a cross-batch scene consistency strategy to capture shared feature information among drone images from the same region. This architecture allows DAC method to better integrate fine-grained structural information with coarse-grained global information, achieving more distinctive and robust feature representations in a multi-grained pattern. Our experimental results demonstrate that the proposed DAC method achieves state-of-the-art performance on three widely used public datasets (University-1652, SUES-200 and Multi-weather University-1652). Moreover, DAC exhibits superior cross-region transferability compared to existing state-of-the-art methods by a large margin. These results showcase the effectiveness of our proposed method in terms of feature representation and cross-view geo-localization performance. In the future, we plan to extend our research to address geo-localization challenges with larger viewpoint differences, such as the alignment between drone/satellite images and ground panoramic images, and explore higher-precision localization tasks, such as 3-DoF geo-localization.

REFERENCES

[1] Z. Zheng, Y. Wei, and Y. Yang, "University-1652: A multi-view multi-source benchmark for drone-based geo-localization," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1395–1403.

[2] S. Brar, R. Rabbat, V. Raithatha, G. Runcie, and A. Yu, "Drones for deliveries," Sutardja Cent. Entrepreneurship Technol., Univ. Calif. Berkeley, Berkeley, CA, USA, Tech. Rep. 8, 2015.

[3] Q. Yu et al., "Building information modeling and classification by visual learning at a city scale," 2019, *arXiv:1910.06391*.

[4] P. Zhu, L. Wen, X. Bian, H. Ling, and Q. Hu, "Vision meets drones: A challenge," 2018, *arXiv:1804.07437*.

[5] Y. Tian, C. Chen, and M. Shah, "Cross-view image matching for geo-localization in urban environments," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 1998–2006, doi: 10.1109/CVPR.2017.216.

[6] T. Wang et al., "Each part matters: Local patterns facilitate cross-view geo-localization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 2, pp. 867–879, Feb. 2022, doi: 10.1109/TCSVT.2021.3061265.

[7] M. Dai, J. Hu, J. Zhuang, and E. Zheng, "A transformer-based feature segmentation and region alignment method for UAV-view geo-localization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 7, pp. 4376–4389, Jul. 2022.

[8] T. Shen, Y. Wei, L. Kang, S. Wan, and Y.-H. Yang, "MCCG: A ConvNeXt-based multiple-classifier method for cross-view geo-localization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 3, pp. 1456–1468, Mar. 2024, doi: 10.1109/TCSVT.2023.3296074.

[9] F. Deuser, K. Habel, and N. Oswald, "Sample4Geo: Hard negative sampling for cross-view geo-localisation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 16847–16856.

[10] Y. Shi and H. Li, "Beyond cross-view image retrieval: Highly accurate vehicle localization using satellite image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 16989–16999, doi: 10.1109/CVPR52688.2022.01650.

[11] T. Lentsch, Z. Xia, H. Caesar, and J. F. P. Kooij, "SliceMatch: Geometry-guided aggregation for cross-view pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 17225–17234, doi: 10.1109/CVPR52729.2023.01652.

[12] Z. Xia, O. Booij, and J. F. P. Kooij, "Convolutional cross-view pose estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 5, pp. 3813–3831, May 2024.

[13] T.-Y. Lin, Y. Cui, S. Belongie, and J. Hays, "Learning deep representations for ground-to-aerial geolocalization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5007–5015.

[14] S. Workman, R. Souvenir, and N. Jacobs, "Wide-area image geolocalization with aerial reference imagery," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3961–3969.

[15] L. Liu and H. Li, "Lending orientation to neural networks for cross-view geo-localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5617–5626.

[16] S. Zhu, T. Yang, and C. Chen, "VIGOR: Cross-view image geo-localization beyond one-to-one retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5316–5325.

[17] R. Zhu, L. Yin, M. Yang, F. Wu, Y. Yang, and W. Hu, "SUES-200: A multi-height multi-scene cross-view image benchmark across drone and satellite," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4825–4839, Sep. 2023.

[18] T. Wang, Z. Zheng, Y. Sun, C. Yan, Y. Yang, and T.-S. Chua, "Multiple-environment self-adaptive network for aerial-view geo-localization," *Pattern Recognit.*, vol. 152, Aug. 2024, Art. no. 110363.

[19] F. Castaldo, A. Zamir, R. Angst, F. Palmieri, and S. Savarese, "Semantic cross-view matching," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 1044–1052.

[20] S. Workman and N. Jacobs, "On the location dependence of convolutional neural network features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 70–78.

[21] Y. Shi, L. Liu, X. Yu, and H. Li, "Spatial-aware feature aggregation for image based cross-view geo-localization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 10090–10100.

[22] Y. Shi, D. Campbell, X. Yu, and H. Li, "Geometry-guided street-view panorama synthesis from satellite imagery," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 10009–10022, Dec. 2022.

[23] S. Zhu, M. Shah, and C. Chen, "TransGeo: Transformer is all you need for cross-view image geo-localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 1152–1161, doi: 10.1109/CVPR52688.2022.00123.

[24] T. Wang, Z. Zheng, Z. Zhu, Y. Gao, Y. Yang, and C. Yan, "Learning cross-view geo-localization embeddings via dynamic weighted decorrelation regularization," 2022, *arXiv:2211.05296*.

[25] J. Lin et al., "Joint representation learning and keypoint detection for cross-view geo-localization," *IEEE Trans. Image Process.*, vol. 31, pp. 3780–3792, 2022, doi: 10.1109/TIP.2022.3175601.

[26] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3733–3742.

[27] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[28] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*. Glasgow, U.K.: Springer, Aug. 2020, pp. 776–794.

[29] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9729–9738.

[30] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[31] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," 2020, *arXiv:2003.04297*.

[32] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 22243–22255.

[33] J.-B. Grill et al., "Bootstrap your own latent—A new approach to self-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 21271–21284.

[34] X. Chen and K. He, "Exploring simple Siamese representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15750–15758.

[35] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9640–9649.

[36] M. Caron et al., "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9650–9660.

[37] M. Oquab et al., "DINOv2: Learning robust visual features without supervision," 2023, *arXiv:2304.07193*.

[38] A. Radford et al., "Learning transferable visual models from natural language supervision," 2021, *arXiv:2103.00020*.

[39] L. Huang, S. You, M. Zheng, F. Wang, C. Qian, and T. Yamasaki, "Learning where to learn in cross-view self-supervised learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 14431–14440, doi: 10.1109/CVPR52688.2022.01405.

[40] D. Misra, T. Nalamada, A. U. Arasanipalai, and Q. Hou, "Rotate to attend: Convolutional triplet attention module," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 3139–3148.

[41] J. Zhuang, M. Dai, X. Chen, and E. Zheng, "A faster and more effective cross-view matching method of UAV and satellite images for UAV geolocalization," *Remote Sens.*, vol. 13, no. 19, p. 3979, Oct. 2021.

[42] R. Zhu, M. Yang, L. Yin, F. Wu, and Y. Yang, "UAV's status is worth considering: A fusion representations matching method for geo-localization," *Sensors*, vol. 23, no. 2, p. 720, Jan. 2023.

Panwang Xia (Graduate Student Member, IEEE) received the B.S. degree from Wuhan University, Wuhan, China, in 2022, where he is currently pursuing the Ph.D. degree with the School of Remote Sensing and Information Engineering. His research interests include deep learning, computer vision, artificial intelligence, image representation learning, and their applications.

Yi Wan (Member, IEEE) was born in 1991. He received the B.S. and Ph.D. degrees from Wuhan University, Wuhan, China, in 2013 and 2018, respectively.

He is currently an Associate Professor with Wuhan University. His research interests include digital photogrammetry, computer vision, 3D reconstruction, and change detection in remote sensing imagery.

Zhi Zheng (Member, IEEE) received the B.S. and Ph.D. degrees in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2017 and 2023, respectively.

He is currently a Post-Doctoral Fellow with the Department of Geography and Resource Management, The Chinese University of Hong Kong (CUHK), where he has also been awarded the Research Fellowship Scheme, set to begin in January 2024. He has published over ten research articles, including one ESI highly cited paper and one featured as the frontispiece, both as the first author. His research interests include satellite remote sensing, 3D reconstruction, land use analysis, and geohazards monitoring using deep learning methods.

Yongjun Zhang (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in photogrammetry and remote sensing from Wuhan University (WHU), Wuhan, China, in 1997, 2000, and 2002, respectively.

He is currently a Professor of photogrammetry and remote sensing with the School of Remote Sensing and Information Engineering, WHU. His research interests include aerospace and low-attitude photogrammetry, image matching, combined block adjustment with multisource data sets, integration of LiDAR point clouds and images, and three-dimensional city reconstruction. His research interests include remote sensing image processing and machine learning.

Jiwei Deng was born in 1996. He received the B.S. degree from Wuhan University, Wuhan, China, in 2010. He is currently a Senior Engineer with China Railway Design Group Company Ltd. His research interests include railway aerial photogrammetry and computer vision.