Learning Cross-view Geo-localization Embeddings via Dynamic Weighted Decorrelation Regularization

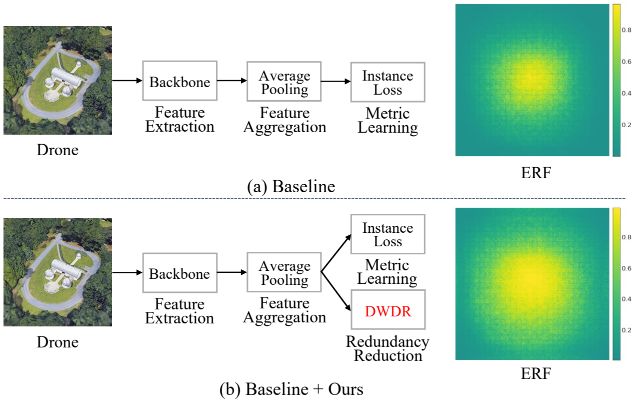
Tingyu Wang, Zhedong Zheng, Zunjie Zhu, Yaoqi Sun, Chenggang Yan, and Yi Yang, *Senior Member, IEEE*

***Abstract*—In the domain of cross-view geo-localization, the challenge lies in accurately matching images captured from distinct perspectives, such as aerial drone imagery and satellite imagery of the same geographical location. Existing methods predominantly concentrate on minimizing distances between feature embeddings in the representational space, inadvertently overlooking the significance of reducing embedding redundancy. This oversight potentially hampers the extraction of diverse and distinctive visual patterns critical for precise localization. This work argues that minimizing embedding redundancy is a pivotal factor in enhancing a model's ability to discriminate diverse scene characteristics. To support this claim, we introduce a straightforward yet effective regularization technique, termed Dynamic Weighted Decorrelation Regularization (DWDR). DWDR serves to actively promote the learning of orthogonal feature channels within neural networks. By dynamically adjusting weights, DWDR targets the minimization of inter-channel correlations, guiding the correlation matrix towards diagonality, indicative of independence among channels. The dynamic weighting mechanism adaptively prioritizes the decorrelation of channels that remain highly correlated throughout training. Additionally, we devise a symmetrical sampling strategy for cross-view scenarios to ensure that the training examples are balanced across different imaging platforms in a batch. Despite its simplicity, the integration of DWDR and the proposed sampling scheme yields remarkable performance across four extensive benchmark datasets: University-1652, CVUSA, CVACT, and VIGOR. Notably, in stringent conditions, such as when constrained to exceedingly compact feature dimensions of 64, our methodology significantly outperforms conventional baselines, thereby affirming its efficacy and robustness under challenging constraints.**

***Index Terms*—Geo-localization, Image Retrieval, Deep Learning, The Cross-correlation Correlation Matrix, Decorrelation.**

I. Introduction

CROSS-VIEW geo-localization is an image retrieval task and has been broadly applied to event detection, drone navigation, and accuracy delivery [1], [4], [5], [6]. Given a probe from the query platform (*e.g.*, the drone), the system aims to spot a candidate image in the gallery platform (*e.g.*, the satellite) containing the same geographic target with the probe. Since satellite images possess GPS metadata as annotations, we can easily acquire the location information of the interesting probe. In addition, when the GPS signal of the positioning device encounters interference, the image-based cross-view geo-localization can also be employed as an auxiliary tool to refine the geo-localization and provide a more robust result.



(a) Baseline

(b) Baseline + Ours

Fig. 1. (a) A strong metric learning based baseline [1]. (b) The baseline combined with our proposed dynamic weighted decorrelation regularization (DWDR). ERF refers to the effective receptive field (the yellow area). A large ERF reflects that the learned model is able to extract more discriminative visual features from a wider range without over-fitting to the local area [2], [3].

Cross-view geo-localization remains challenging since images from different platforms inherently contain viewpoint variations. The extreme viewpoint change leads to differences in the visual appearance of a geographic target, which confuses the system to locate a position accurately. A crucial key to the geo-localization challenge is to learn a discriminative visual embedding [7], [8], [9], [10], [11]. Recently, deep learning technologies have received much research attention in the cross-view geo-localization problem since the great potential in feature extraction. A popular scheme for learning cross-view geo-localization models is first utilizing the pre-trained convolutional neural network to extract feature maps of images. Following, various metric learning functions are proposed to pull the image pairs with the same geo-tag closer while pushing those features from non-matchable pairs far apart [1], [12], [13], [14]. Based on this basic scheme, the attention mechanism [15], [16] and aligning the spatial layout of features [4], [17], [18], [19] are also widely considered in the network design. Most existing methods focus more on the similarity between cross-view embeddings while ignoring the

Tingyu Wang, and Chenggang Yan are with the School of Communication Engineering, Hangzhou Dianzi University, China 310018 (e-mail: tingyu.wang@hdu.edu.cn; cgyan@hdu.edu.cn).

Zhedong Zheng is with the Faculty of Science and Technology, and Institute of Collaborative Innovation, University of Macau, China (e-mail: zhedongzheng@um.edu.mo).

Zunjie Zhu and Yaoqi Sun are with the School of Communication Engineering, Hangzhou Dianzi University, China 310018, also with the Lishui Institute of Hangzhou Dianzi University, China 323000 (e-mail:zunjiezhu@hdu.edu.cn; syq@hdu.edu.cn). Yaoqi Sun is the Corresponding Author.

Yi Yang is with the School of Computer Science, Zhejiang University, China, 310027 (e-mail: yangyics@zju.edu.cn).

redundant channels of the embedding itself.

In human perception study, the neuroscience H. Barlow claims that concise, non-redundant descriptions are of higher value to the perception system and will help to clarify inputs of the external world [20]. Based on this bio-perceptual hypothesis, we argue that stripping the redundancy of visual embeddings contributes to the discrimination of different targets. In this paper, we propose a dynamic weighted decorrelation regularization (DWDR). The measuring objective of DWDR is the Pearson cross-correlation coefficient matrix, which is computing from features of positive pairs composed of cross-view images. Specifically, DWDR employs Square Loss to regress the diagonal elements of the objective matrix to 1, and the off-diagonal elements are approximated to 0. However, the objective matrix is typically large, *e.g.*, 2048×2048. The optimizer is easily overwhelmed by a mass of elements already close to the optimization goal, thus ignoring other elements that need to be regressed. To address this optimization problem, we assign a dynamic weight to each element loss according to the maximum regression distance of the target element. The dynamic weight can adaptively highlight the importance of poorly-regressed elements and suppress the side effect of well-regressed elements in optimization. We observe that DWDR encourages the learned model to focus on a larger effective receptive field, which prevents the network from overfitting to a local pattern [2] (see Figure 1). Besides, as mentioned above, the computation of the Pearson cross-correlation coefficient matrix requires positive sample pairs. As a by-product of selecting positive pairs, we also provide a cross-view symmetric sampling strategy. In a training batch, our symmetric sampling strategy aligns the number of the same geo-tag images between different platforms. Therefore, the proposed strategy mitigates the sample imbalance, especially in drone-to-satellite geo-localization, which contains limited satellite data [21], [22]. To summarize, our contributions are as follows.

- We propose a dynamic weighted decorrelation regularization (DWDR), which motivates networks to learn discriminative embeddings by stripping the redundancy of features. During training, DWDR assigns a dynamic weight to the loss of each element in the objective matrix, yielding efficient optimization of networks. As a by-product of DWDR, we further introduce a cross-view symmetric sampling strategy, which maintains the example balance in a training batch.
- Albeit simple, we demonstrate the effectiveness of the proposed method on four cross-view geo-localization datasets. Extensive experiments show that multiple existing works [1], [4], [23] fused with our method are able to further boost the performance. In addition, we observe that our method still obtains superior results even for a short visual embedding with 64 dimensions.

The rest of this paper is organized as follows. In Section II, we discuss related works. The details of our method are illustrated in Section III. Experimental results are provided in Section IV. Finally, Section V presents a summary.

II. Related Work

In this section, we briefly review related previous works, including image-based cross-view geo-localization and low-redundancy representation learning.

A. Image-based Cross-view Geo-localization

Imaged-based cross-view geo-localization has been tackled as an image retrieval task. Early works [24], [25], [26] employ hand-crafted operators to extract discriminative features for cross-view image matching. With the development of deep learning, the convolutional neural network (CNN) has received much research attention in the extraction of image representation. The pioneering CNN-based approach [27] directly deploys pre-trained AlexNet [28] to extract features for the cross-view geo-localization. Further, [29] introduce the information of image pairs as the constraint to fine-tune the pre-trained network and acquire a better performance. Following this line of considering object constraints, [30] borrow knowledge from face verification and harness the contrastive loss [31] to guide the optimization of a modified Siamese Network [32]. [33] discuss the limitation of the Siamese architecture in large-scale cross-view matching and provide a soft-margin triplet loss to improve the geo-localization accuracy. Similarly, [34] propose a weighted soft-margin ranking loss, which not only improves the matching accuracy but also speeds up the training convergence. [35] mine hard examples in the training batch to strengthen the penalization of the soft-margin triplet loss. [1] suggest that images with the same identification can be classified into one cluster and apply the instance loss [36], [37], [38] as the proxy target to learn discriminative embeddings. Another line of works concentrates on addressing the spatial misalignment problem of cross-view retrieval. CVM-Net [34] employs a shared NetVLAD to aggregate the local feature to minor the visual gap between different viewpoints. [17] explicitly encode the orientation information into the image descriptors and boost the discriminative power of the learned features. [18] first attempt to utilize the optimal transport (OT) theory to close the spatial layout information in the high-level feature. Then [16] directly resort to the polar transform to align the pixel-level semantic information of cross-view images. DSM [39] designs a dynamic similarity matching module to solve the cross-view matching in a limited Field of View (FoV). LPN [4] stresses the importance of contextual information and proposes a square-ring partition strategy to improve the performance of cross-view geo-localization. Without any extra annotations, RK-Net [15] automatically detects salient keypoints to improve the model capability against the appearance changes.

B. Low-Redundancy Representation Learning

In the early study of human perception, the neuroscientist H. Barlow [20] suggests that the perception system tends to encode the raw sensory input as the low-redundancy representation in which each component possesses statistical independence. This learning principle guides a number of algorithms in machine learning. [40] support that the decorrelation criterion is useful in the context of data and derive a fast online pretraining algorithm to learn decorrelated features for neural networks. [41] design Decov Loss that motivates the network to learn non-redundant representations and demonstrate that decorrelating representations helps to reduce overfitting of the trained deep networks. [42] utilize Singular Value Decomposition (SVD) and reduce the correlation between output units by integrating the orthogonality constraint in CNN training. Thus the final descriptor contains lower redundant information about the sample. In self-supervised learning (SSL), Barlow Twins [43] proposes a simple yet effective object function to acquire representations with low redundancy and avoid model collapse. The optimization goal of Barlow Twins is to transform a cross-correlation matrix into an identity matrix. SSL does not require the input data with human annotation, and the cross-correlation matrix is computed from two distorted representations of a sample.

III. Proposed Method

In Section III-A, we first give a revisit of preliminaries followed by the description of our baseline network structure for geo-localization in Section III-B. Next, we introduce the dynamic weighted decorrelation regularization (DWDR). The dynamic weight mechanism relieves the plateau problem in Barlow Twins [43]. We also provide a mechanism discussion (see Section III-C).

A. Preliminaries

Cross-correlation matrix measures the correlation between two matrices. In particular, for two random matrices $\mathbf{X} = (X_1, X_2, \cdots, X_M)^T$, $\mathbf{Y} = (Y_1, Y_2 \cdots, Y_N)^T$, where $X_m, Y_n \in \mathbb{R}^a$ with a dimensions, the cross-correlation matrix of $\mathbf{X}$ and $\mathbf{Y}$ can be defined as:

$$\phi \triangleq \mathbb{E}\left[\mathbf{X}\mathbf{Y}^{\mathbf{T}}\right]. \tag{1}$$

A component-wise description is:

$$\phi = \begin{bmatrix} \mathbb{E}\left[X_1 Y_1^T\right] & \mathbb{E}\left[X_1 Y_2^T\right] & \cdots & \mathbb{E}\left[X_1 Y_N^T\right] \\ \mathbb{E}\left[X_2 Y_1^T\right] & \mathbb{E}\left[X_2 Y_2^T\right] & \cdots & \mathbb{E}\left[X_2 Y_N^T\right] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}\left[X_M Y_1^T\right] & \mathbb{E}\left[X_M Y_2^T\right] & \cdots & \mathbb{E}\left[X_M Y_N^T\right] \end{bmatrix}, \tag{2}$$

where $\mathbb{E}[\cdot]$ refers to the expectation.

Pearson cross-correlation coefficient matrix is similar to the cross-correlation matrix, and it shows a normalized measurement between two matrices. Differently, the value of each element in the Pearson cross-correlation coefficient matrix is between -1 and 1. The score 1 and -1 denote that two vectors are perfectly correlated and anti-correlated. 0 means that two vectors are completely unrelated. Based on the definition of the cross-correlation matrix, the Pearson cross-correlation coefficient matrix can be formulated as:

$$\rho = \frac{\mathbb{E}\left[(\mathbf{X} - \mu_{\mathbf{X}})(\mathbf{Y} - \mu_{\mathbf{Y}})^T\right]}{\sigma_{\mathbf{X}}\sigma_{\mathbf{Y}}}, \tag{3}$$

where $\mu_{\mathbf{X}}$ and $\sigma_{\mathbf{X}}$ are the mean and the standard deviation of $\mathbf{X}$, respectively. $\mu_{\mathbf{Y}}$ and $\sigma_{\mathbf{Y}}$ are the mean and the standard deviation of $\mathbf{Y}$, separately. $\mathbb{E}[\cdot]$ is the expectation.

Barlow Twins [43] is a widely-used method for self-supervised learning (SSL). To address the issue of representation collapse, the main contribution of Barlow Twins is introducing a new regularization objective, which can be defined as:

$$L_{BT} \triangleq \sum_{i=1}^{d}(1 - \phi_{ii})^2 + \lambda \sum_{i=1}^{d} \sum_{j=1, j\neq i}^{d} {\phi_{ij}}^2, \tag{4}$$

where λ is a positive hyper-parameter, and ϕ is the cross-correlation matrix between two mini-batch of features. Given two batches of features with d dimensions (d is generally set as 2048), we multiply the feature matrix along with the batch dimension. Thus, the dimension of ϕ is $d \times d$. The regularization function L_{BT} encourages every feature channel to be independent of others. Specifically, it impels the diagonal elements ϕ_{ii} from the same channel to 1, while pushing off-diagonal elements ϕ_{ij} between different channels to 0. However, in practice, Barlow Twins meets the optimization problem, especially when facing a typical large matrix (*e.g.*, 2048 × 2048). The model arrives at the plateau after the majority of channels are converged, and it neglects other still-correlated "hard" channels, compromising the training process.

B. Network Structure

We adopt ResNet-50 [44] as the backbone and add a new classifier. The classifier consists of a fully-connected layer (FC), a batch normalization layer (BN), a dropout layer (Dropout), and another fully-connected layer (FC). Notably, the backbone can also be other networks such as VGG16 [45] and Swin Transformer [23]. The two-branch baseline consists of three forward phases, *i.e.*, feature extraction, feature aggregation, and feature classification. Specifically, we denote the input images from two platforms as x_k, where $k \in \{1, 2\}$. 1 denotes the satellite platform, and 2 refers to the drone or ground platform. We first employ two backbones with shared weights to extract feature maps. Then the global average pooling is deployed to aggregate the information of feature maps into the column vectors f_k. Finally, we harness a classifier to map vectors f_k of different platforms into one shared classification space and acquire the predicted logit vectors z_k. Meanwhile, the cross entropy function is employed to calculate the instance loss L_{id} [1]. The instance loss is a classification loss with a shared classifier $\mathcal{F}_{classifier}$. The above process can be formulated as:

$$f_k = Avgpool(\mathcal{F}^k_{backbone}(x_k)), \tag{5}$$

$$z_k = \mathcal{F}_{classifier}(f_k), \tag{6}$$

$$L_{id} = \sum_{k=1}^{2} -log \frac{exp(z_k(y))}{\sum_{c=1}^{C} exp(z_k(c))}. \tag{7}$$

The label $y \in [1, C]$, where C indicates the category number of geographic targets in the training set. $z_k(y)$ is the logit score

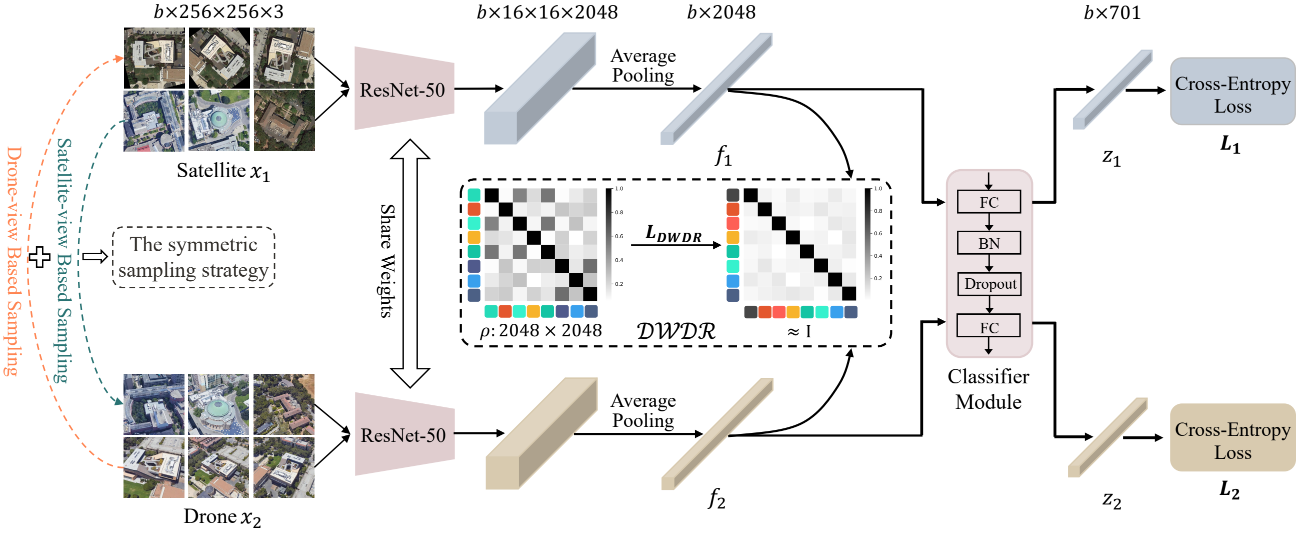


Fig. 2. A schematic overview of our method. We first apply the symmetric sampling strategy to generate one batch of drone-and-satellite positive pairs. The symmetric sampling strategy is composed of the drone-view based sampling and the satellite-view based sampling. Then the satellite-platform and the drone-platform images of the batch are fed into a two-branch network. The two-branch network shares weights of two backbones since images from the drone platform and the satellite platform have similar patterns. Next, the average pooling is deployed to aggregate the output feature maps of each branch into the column vectors. Finally, the column vectors of two branches are inputted into a classifier module to acquire predicted logit scores, respectively, and the cross-entropy function is utilized to compute the instance loss. The proposed DWDR aims to transform the Pearson cross-correlation coefficient matrix ρ into an identity matrix I as much as possible. The Pearson cross-correlation coefficient matrix ρ is calculated by column vectors from two branches. Note that here we show the framework employing ResNet-50 as the backbone and images of University-1652 as inputs. When training on CVUSA, CVACT and VIGOR, the two-branch backbone does not share weights.

of the ground-truth geo-tag y. When inference, we remove the final linear classification layer and extract the intermediate feature f_k as the visual representation.

C. Dynamic Weighted Decorrelation Regularization

In this work, we introduce a dynamic weighted decorrelation regularization (DWDR) to encourage the network to learn low-redundancy visual embeddings. As shown in Figure 2, DWDR is implemented based on a classic two-branch baseline [1]. The two-branch baseline harnesses location classification as the pretext [1], [46] to conduct the cross-view geo-localization task. During training, we employ the symmetric sampling strategy to balance examples between different platforms in a training batch. It is worth noting that the symmetric sampling strategy is a by-product of DWDR.

The optimization objective of DWDR is the Pearson cross-correlation coefficient matrix ρ between f_k extracted from images of different platforms. Given two batches of extracted vectors f_1 and f_2 of size $b \times 2048$, according to Eq. 3, we can gain the objective matrix ρ with the shape of 2048×2048, where b denotes the batchsize. DWDR aims to spur the network by regressing the objective matrix ρ into a sparse matrix, *i.e.*, an identity matrix. We employ Square Loss to constrain the regression of each element. DWDR can be written as:

$$L_{DWDR} \triangleq \sum_{i=1}^{d} \omega_1 \cdot (1-\rho_{ii})^2 + \lambda \sum_{i=1}^{d} \sum_{j=1, j\neq i}^{d} \omega_2 \cdot \rho_{ij}^2, \tag{8}$$

where λ is a hyper-parameter to balance the diagonal and off-diagonal element weight, ρ_{ii} refers to the diagonal elements of the objective Pearson matrix ρ, and ρ_{ij} denotes off-diagonal elements. ρ_{ii} is regressed to 1, which makes visual embeddings of the same geo-tag invariant for different platforms. ρ_{ij} is regressed to 0 to make the visual embedding channels independent from each other. ω_1 and ω_2 are two dynamic weights to prevent the optimization plateau, which depends on the regression score. In this way, the dynamic weight adjusts the influence of the loss and adaptively pays attention to the poorly-regressed elements during training. Considering that each element of the Pearson matrix is in $[-1, 1]$, we set $\omega_1 = \left(\frac{1-\rho_{ii}}{2}\right)^{\gamma_1}, \omega_2 = |\rho_{ij}|^{\gamma_2}$, $|\cdot|$ denotes the absolute value. In this way, given non-negative focusing parameters γ_1 and γ_2, we ensure $\omega_1 \in [0, 1]$ and $\omega_2 \in [0, 1]$. For elements close to the regression result (**well-regressed elements**), the assigned dynamic weight is near 0. Conversely, for elements far from the regression target (**poorly-regressed elements**), the assigned dynamic weight increases to 1.

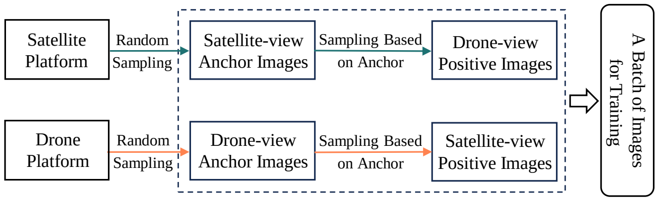


Fig. 3. A diagram of the symmetric sampling strategy.

Symmetric sampling strategy. In order to compute the Pearson cross-correlation coefficient matrix, a necessary step is to construct a training batch by acquiring different-platform images with same geo-tags ('positive pairs'). Here we take the University-1652 dataset [1] as an example. A satellite-view image is randomly selected as an anchor to create a positive pair with the corresponding drone-view image that shares the same geo-tag, a process termed satellite-view based sampling. Conversely, when the drone-view image serves as the anchor, the method is referred to as drone-view based sampling. We note that the image number of different platforms is usually different due to the capturing difficulty. For example, we can easily acquire multiple drone images while only having one satellite image. If we apply the satellite-view based sampling during training, it will miss sampling some drone-view images. It is because every time we randomly sample one image from 54 drone-view images with putting back. On the other hand, if we apply the drone-view based sampling, it will contain duplicate geo-tags within a mini-batch, which repeatedly samples the same satellite image. Therefore, we propose a symmetric sampling strategy (see Figure 3), which combines the satellite-view based sampling and drone-view based sampling. In particular, we sample two mini-batch by the two strategies respectively and combine them together as the new mini-batch to train the model. This combined strategy ensures the model can "see" all the training data while keeping the category sampling relatively balanced.

Discussion. The proposed method is similar to Barlow Twins considering disentangling the correlation matrix, but is different in the following aspects: First, Barlow Twins [43] optimizes the cross-correlation matrix, while our method harnesses the Pearson cross-correlation coefficient matrix. The Pearson matrix is preferable, since it normalizes the element in a limited range of [-1,1], which unifies the element scale within the matrix and prevents overflowing. Second, as shown in Eq. 4, Barlow Twins accumulated the error along the whole matrix ϕ. However, the dimension of ϕ is large, *e.g.*, 2048 × 2048. As training proceeds, the majority of elements converge, and the network arrives at the plateau, since the loss is accumulated by a vast of elements. The optimization of the remaining minority elements is usually ignored. The proposed DWDR also accumulated the error but with dynamic weights for different elements in Eq. 8. We leverage the Pearson matrix, which is normalized in the range [-1, 1], to set the corresponding dynamic weights. The design also limits dynamic weights in [0,1], preventing the weight overflow. Therefore, DWDR can focus on the minority elements, even when majority channels are converged. Compared with Barlow Twins, DWDR encourages the network to make still correlated channels independent throughout the training period.

Optimization. We optimize our network by jointly employing the instance loss and DWDR:

$$L_{total} = \alpha L_{id} + (1-\alpha) L_{DWDR}. \tag{9}$$

The instance loss L_{id} forces different-platform images with the same geo-tag to be close on the high-level features and pushes mismatched images far apart. At the same time, DWDR motivates the learned visual embeddings with independent channels. Thus the network is able to extract more discriminative features. α is a weight to control the importance of the loss function and the regularization term.

IV. Experiment

We introduce four cross-view geo-localization datasets and the evaluation protocol in Section IV-A. The implementation detail is provided in Section IV-B. We carry out a series of comparisons with state-of-the-art approaches in Section IV-C, followed by ablation studies in Section IV-D. Finally, Section IV-E visualizes the cross-view geo-localization results.

A. Datasets and Evaluation Protocol

We conduct experiments on four geo-localization datasets, *i.e.*, University-1652 [1], CVUSA [47], CVACT [17] and VIGOR [48].

University-1652 [1] is a multi-view multi-source dataset, including data from three different platforms, *i.e.*, drones, satellites and dash cams. As the name implies, this dataset collects 1652 ordinary buildings of 72 universities around the world. 701 of all 1652 buildings are separated into the training set, and the other 951 builds constitute the testing set. Therefore, build images in the training and testing set are not overlapping. For each building, the dataset contains one satellite-view image, 54 drone-view images and 3.38 ground-view images on average. Since dash cams are hard to acquire enough street-view images for some buildings, the dataset also collects 21,099 common-view images from Google Image as an extra training set. The dataset supports two new aerial-view geo-localization tasks, *i.e.*, drone-view target localization (Drone → Satellite) and drone navigation (Satellite → Drone).

CVUSA [47] is a large-scale cross-view dataset, which consists of images from two viewpoints, *i.e.*, the ground view and the satellite view. In the dataset, 35,532 ground-and-satellite image pairs are employed for training, and 8,884 image pairs are provided for testing. Noteworthily, ground-view images are the pattern of panoramas, and the orientation of all ground and satellite images is aligned.

CVACT [17] is a similar dataset to CVUSA. For the ground-to-satellite task, CVACT also contains 35,532 image pairs for training. Different from CVUSA, CVACT provides a validation set with 8,884 image pairs denoted as CVACT_val. Meanwhile, compared to CVUSA, CVACT possesses a larger test set with 92,802 image pairs named CVACT_test. When evaluated in CVACT_val, a query ground-view panorama matches only one satellite image in the gallery. However, for CVACT_test, a panoramic query image may correspond to several satellite images within 5 meters from the ground-truth location.

VIGOR [48] considers the misalignment in spatial location of ground-to-satellite cross-view geo-localization. Different with one-to-one retrieval, VIGOR consists of 90,618 satellite images and 105,214 ground panoramas, and one satellite may against two ground panoramas. In evaluation, VIGOR supports two application scenarios, *i.e.*, same-area and cross-area protocols. Satellite images in the same-area protocol are invariant when training and testing, but are different when executing the cross-area protocol.

We follow existing works [4], [18], [16] and mainly employ CVACT_val to evaluate our method when training on CVACT. Besides, we only apply the symmetric sampling strategy in

TABLE I
STATISTICS OF FOUR CROSS-VIEW DATASETS, INCLUDING THE IMAGE NUMBER OF TRAINING AND TESTING SETS. THE LEFT AND RIGHT OF THE ARROW (→) REFER TO THE QUERY AND GALLERY PLATFORMS, RESPECTIVELY.

Dataset	Training		Testing			
	Drone	Satellite	Drone →	Satellite	Satellite →	Drone
University-1652 [1]	37,854	701	37,855	951	701	51,355
	Ground	Satellite	Ground →	Satellite	Satellite →	Ground
CVUSA [47]	35,532	35,532	8,884	8,884	8,884	8,884
CVACT [17]	35,532	35,532	8,884	8,884	8,884	8,884
VIGOR [48](Same-Area)	52,609	90,618	52,605	90,618	90,618	52,605
VIGOR [48](Cross-Area)	51,520	44,055	53,694	46,563	46,563	53,694

TABLE II
COMPARISON WITH EXISTING RESULTS REPORTED ON UNIVERSITY-1652. THE COMPARED METHOD ARE CATEGORIZED INTO THREE GROUPS. THE FIRST GROUP CONSISTS OF BASELINE-RELATED METHODS WHICH EMPLOY AVERAGE POOLING TO AGGREGATE FEATURE MAPS. THE SECOND GROUP CONTAINS METHODS THAT APPLY CONTEXTUAL INFORMATION. THE THIRD GROUP INCLUDES TRANSFORMER-BASED METHODS. "M" INDICATES THE MARGIN OF THE TRIPLET LOSS. † DENOTES THE INPUT IMAGE OF SIZE 384×384. THE INPUT IMAGE SIZE OF TWO TRANSFORMER-BASED METHODS AND OTHER CNN-BASED METHODS ARE 224×224 AND 256×256, RESPECTIVELY. THE BEST RESULTS ARE IN BOLD.

Method	University-1652			
	Drone → Satellite		Satellite → Drone	
	R@1	AP	R@1	AP
Instance Loss (**Baseline**) [1]	57.09	61.88	73.89	58.73
Contrastive Loss [30]	52.39	57.44	63.91	52.24
Triplet Loss ($M = 0.3$) [49]	52.16	57.47	65.05	52.37
Triplet Loss ($M = 0.5$) [49]	51.23	56.40	62.77	51.29
Soft Margin Triplet Loss [34]	53.67	58.69	67.90	54.76
LCM† [21]	66.65	70.82	79.89	65.38
RK-Net [15]	66.13	70.23	80.17	65.76
Baseline [1] + Ours	**69.77**	**73.73**	**81.46**	**70.45**
LPN [4]	75.93	79.14	86.45	74.49
LPN + USAM [15]	77.60	80.55	86.59	75.96
PCL [50]	79.47	83.63	87.69	78.51
LPN [4] + Ours	**81.51**	**84.11**	**88.30**	**79.38**
Swin-B [23]	84.15	86.62	90.30	83.55
FSRA [22]	84.51	86.71	88.45	83.47
Swin-B [23] + Ours	**86.41**	**88.41**	**91.30**	**86.02**

University-1652 because there is an obvious imbalance of training examples across different platforms, *i.e.*, 1 satellite image corresponds to 54 drone images (see Table I).

Evaluation protocol. In our experiments, the performance of our method is evaluated by three metrics, *i.e.*, Recall@K (**R@K**), the average precision (**AP**) and the hit rate. **R@K** refers to the proportion of true-matched candidates in the top-K of the ranking list. The value of **AP** is measured by the area under the Precision-Recall curve. The hit rate refers to the ratio of correctly matched top-1 reference images to query images. Higher scores of these three metrics denote better network performance.

B. Implementation Details

Our method is performed based on a classic two-branch baseline [1]. The baseline adopts a modified ResNet-50 [44] pre-trained on ImageNet [51] to extract visual features. Specifically, we remove the final classification layer of ResNet-50 [44]. Besides, the stride of the second convolution layer and the down-sample layer in the first bottleneck of the ResNet-50 [44] stage4 is set from 2 to 1. The input image is resized to 256×256, and the image augmentation consists of random cropping, random horizontal flipping, and random rotation. We employ stochastic gradient descent (SGD) with momentum 0.9 and weight decay 0.0005 to update model parameters. The image number of each platform in a mini-batch is 16. The initial learning rate is 0.001 for the modified ResNet-50 [44] backbone and 0.01 for the classifier module. The dropout rate in the classifier module is 0.75. In all datasets, we train our models for 120 epochs, and the learning rate is decayed by 0.1 after 80 epochs. Also, except for the proposed DWDR, we only adopt the instance loss for model optimization. There are two trade-off parameters λ and α in the loss function. We run a simple search and observe the better results for $\lambda = 1.3 \times 10^{-3}$ and $\alpha = 0.9$. Note that when using Swin-B [23] and VGG16 [45] as backbones, $\lambda = 2.0 \times 10^{-3}$ and $\lambda = 3.9 \times 10^{-3}$ are best choices, separately. During testing, we deploy the Euclidean distance to compute the similarities between the query and candidates. Our model is implemented on Pytorch [52], and all experiments are conducted on a single NVIDIA RTX 2080Ti GPU.

C. Comparison with Competitive Methods

Results on University-1652. As shown in Table II, we compare our method with lots of competitive methods on University-1652. The compared methods are divided into three groups, *i.e.*, baseline-related methods, methods harnessing contextual information and Transformer-based methods. In the first group, the first line reports results of our two-branch baseline, *i.e.*, "Instance Loss [1]". We can observe that "Baseline + Ours" substantially improves the baseline performance. In the drone-view target localization task (Drone → Satellite), the accuracy of R@1 increases from 57.09% to 69.77% (+12.68%), and the value of AP raises from 61.88% to 73.73% (+11.85%). In the drone navigation task (Satellite → Drone), the accuracy of R@1 goes up from 73.89% to 81.46% (+7.57%), and the value of AP increases from 58.73% to 70.45% (+11.72%). Meanwhile, the performance of our method also has surpassed other baseline-related methods. In the second group, LPN [4] explicitly takes advantage of contextual information during training. Some methods, *e.g.*, "LPN + USAM [15]" and PCL [50], combined with LPN have yielded better results, and we can also implement our method based on LPN. Specifically, we re-implement LPN by utilizing the symmetric sampling strategy to replace the original random sampling and incorporating the dynamic weighted decorrelation regularization during training. Compared with results of LPN, "LPN + Ours" achieves 81.51% R@1 accuracy (+5.58%) and 84.11% AP (+4.97%) on Drone → Satellite and 88.30% R@1 accuracy (+1.85%) and 79.38% AP (+4.89%) on Satellite → Drone. The feature expression ability of Transformer is stronger than that of CNN, and both Transformer-based methods [23], [22] obtain a better

TABLE III

COMPARISON WITH PRIOR ART ON CVUSA AND CVACT. THE COMPARED METHODS ARE DIVIDED INTO 2 COLUMNS. COLUMN1: METHODS WITHOUT THE POLAR TRANSFORM. COLUMN2: METHODS UTILIZING THE POLAR TRANSFORM. "POLAR TRANSFORM" IS THE BOUNDARY OF TWO GROUP COLUMNS. ‡: THE METHOD IS IMPLEMENTED USING IMAGES PROCESSED BY THE POLAR TRANSFORM. ⋆: THE METHOD HARNESSES EXTRA ORIENTATION INFORMATION AS INPUT. THE BEST RESULTS ARE IN BOLD.

Method	Publication	Backbone	CVUSA				CVACT_val			
			R@1	R@5	R@10	R@Top1%	R@1	R@5	R@10	R@Top1%
MCVPlaces [29]	ICCV'15	AlexNet	-	-	-	34.40	-	-	-	-
Zhai [47]	CVPR'17	VGG16	-	-	-	43.20	-	-	-	-
Vo [33]	ECCV'16	AlexNet	-	-	-	63.70	-	-	-	-
CVM-Net [34]	CVPR'18	VGG16	18.80	44.42	57.47	91.54	20.15	45.00	56.87	87.57
Orientation⋆ [17]	CVPR'19	VGG16	27.15	54.66	67.54	93.91	46.96	68.28	75.48	92.04
Zheng (**Baseline**) [1]	MM'20	VGG16	43.91	66.38	74.58	91.78	31.20	53.64	63.00	85.27
Regmi [53]	ICCV'19	X-Fork	48.75	-	81.27	95.98	-	-	-	-
RKNet [15]	TIP'22	USAM	52.50	-	-	96.52	40.53	-	-	89.12
Siam-FCANet [35]	ICCV'19	ResNet-34	-	-	-	98.30	-	-	-	-
CVFT [18]	AAAI'20	VGG16	61.43	84.69	90.94	99.02	61.05	81.33	86.52	95.93
LPN [4]	TCSVT'21	ResNet-50	85.79	95.38	96.98	99.41	79.99	90.63	92.56	97.03
LPN + USAM [15]	TIP'22	ResNet-50	91.22	-	-	99.67	82.02	-	-	**98.18**
Polar Transform										
SAFA [16]	NIPS'19	VGG16	89.84	96.93	98.14	99.64	81.03	92.80	94.84	98.17
DSM [54]	CVPR'20	VGG16	91.96	97.50	98.54	99.67	82.49	92.44	93.99	97.32
4SCIG [55]	TGRS'24	VGG16	92.91	98.15	98.99	99.79	83.18	**93.35**	**95.16**	**99.30**
LPN‡ [4]	TCSVT'21	ResNet-50	93.78	98.50	99.03	99.72	82.87	92.26	94.09	97.77
Baseline + Ours	-	VGG16	75.62	90.45	93.60	98.60	66.76	83.34	87.11	95.10
LPN‡ [4] + Ours	-	ResNet-50	**94.33**	**98.54**	**99.09**	**99.80**	**83.73**	92.78	94.53	97.78

performance than CNN-based methods. We further combine our method with "Swin-B [23]". "Swin-B" indicates the two-branch baseline applying Swin-B as the backbone. "Swin-B + Ours" on University-1652 achieves the state-of-the-art results, *i.e.*, 86.41% in R@1 accuracy and 88.41% in AP for Drone → Satellite and 91.30% in R@1 accuracy and 86.02% in AP for Satellite → Drone.

Results on CVUSA and CVACT. Comparisons with other competitive approaches on CVUSA and CVACT are summarized in Table III. CVUSA and CVACT have a similar data pattern, *i.e.*, the satellite-platform images of aerial viewpoint and the ground panoramas. The polar transform considers the geometric correspondence of two-platform images and transforms the aerial-view image to approximately align a ground panorama at the pixel level. The aligned images help to improve the performance of models. Depending on whether or not the polar transform is harnessed, the compared method can be divided into two columns. The first column reports methods without using polar transform, and methods in the second column employ the polar transform during training and testing. Our method does not employ the polar transform. Experiments on CVUSA and CVACT show phenomena similar to that on University-1652. Our method first outperforms a dual-stream baseline (*i.e.*, the method of Zheng [1]) by a large margin, *i.e.*, 31.71% R@1 improvement on CVUSA and 35.56% R@1 raising on CVACT. At the same time, our method exceeds most of existing methods in the first column. In particular, our method obtains 75.62% R@1, 90.45% R@5, 93.60% R@10, and 98.60% R@Top1% on CVUSA, and 66.76% R@1, 83.34% R@5, 87.11% R@10, and 95.10% R@Top1% on CVACT. In experiments of University-1652, we observe that our method can combine with LPN [4] and achieve better results. The same experiments are also carried out on CVUSA and CVACT. There are two versions of LPN (*i.e.*, LPN and LPN‡) in Table III. LPN‡ applies the polar transform and has achieved higher performance. We notice that our approach still yields competitive results when complemented with the LPN‡. For instance, "LPN‡ + Ours" boosts the R@1 accuracy from 93.78% to 94.33% on CVUSA and 82.87% to 83.73% on CVACT.

Results on VIGOR. We conduct experiments on VIGOR to further validate the applicability of our method. Table IV presents six previous methods for comparisons with our method. To ensure fairness, we report the accuracy of TransGeo [56] without the generalization technology (i.e. ASAM [57]).The observation indicates that our method can still significantly improve the performance of LPN and pushes LPN to achieve competitive results. Specifically, under the same-area protocols over VIGOR, "LPN+Ours" goes up R@1 from 51.95% to 58.00% (+6.05%) and Hit Rate from 63.74% to 69.38% (+5.64%). At the same time, in the cross-area setting, R@1 and Hit Rate increase by 5.87% and 6.59%, respectively.

The above experimental results on four cross-view geo-localization datasets suggest two points. One is that our method can be flexibly applied in different cross-view settings. The other is that our method is able to encourage existing approaches to mine more diverse patterns, yielding discriminative features.

D. Ablation Studies

To further analyze our method, we design several ablation studies. The ablation studies are mainly based on the drone-view target localization (Drone → Satellite) and drone navigation (Satellite → Drone) of University-1652 [1].

Analysis of parameters γ_1 and γ_2. The main contribution of our paper is the proposed dynamic weighted decorrelation

TABLE IV
COMPARISON WITH OTHER COMPETITIVE METHODS ON VIGOR. THE BEST RESULTS ARE IN BOLD. "SWIN-B" INDICATES EMPLOYING SWIN-B AS THE BACKBONE. "W/O ASAM" MEANS THE METHOD WITHOUT ASAM.

Method	R@1	R@5	Hit Rate
SAME			
Siamese-VGG [58]	18.69	43.64	21.90
SAFA [16]	33.93	58.42	36.87
SAFA+Mining [48]	38.02	62.87	41.81
VIGOR [48]	41.07	65.81	44.71
LPN(Swin-B) [4]	51.95	81.06	63.74
TransGeo (w/o ASAM) [56]	52.65	78.29	59.60
LPN+Ours	**58.00**	**84.47**	**69.38**
CROSS			
Siamese-VGG [58]	2.77	8.61	3.16
SAFA [16]	8.20	19.59	8.85
SAFA+Mining [48]	9.23	21.12	9.92
VIGOR [48]	11.00	23.56	11.64
LPN(Swin-B) [4]	12.62	27.81	13.45
TransGeo (w/o ASAM) [56]	13.30	36.20	14.50
LPN+Ours	**18.49**	**37.51**	**20.04**

TABLE V
ABLATION STUDY WITH DIFFERENT γ_1 AND γ_2 IN THE DYNAMIC WEIGHTED DECORRELATION REGULARIZATION. *BT* REFERS TO BARLOW TWINS [43].

Method	γ_1	γ_2	Drone → Satellite		Satellite → Drone	
			R@1	AP	R@1	AP
Baseline [1]+*BT*	0	0	67.91	71.99	80.17	68.03
Baseline [1]+Ours	0	1	66.83	71.01	77.89	68.01
	1	0	67.57	71.62	78.03	67.93
	1	1	**69.77**	**73.73**	**81.46**	**70.45**
	2	2	69.40	73.33	80.88	70.05
LPN [4]+*BT*	0	0	80.93	83.60	86.02	78.33
LPN [4]+Ours	0	1	80.84	83.50	87.30	79.26
	1	0	80.83	83.49	88.30	**79.93**
	1	1	**81.51**	**84.11**	88.30	79.38
	2	2	80.49	83.17	**88.45**	79.91
Swin-B [23]+*BT*	0	0	86.03	88.05	91.01	85.07
Swin-B [23]+Ours	0	1	85.94	88.00	91.01	85.33
	1	0	86.07	88.09	90.30	85.68
	1	1	**86.41**	**88.41**	**91.30**	**86.02**
	2	2	85.54	87.73	90.58	85.65

TABLE VI
THE EXPLORATIONS OF DWDR ABOUT OPTIMIZING DIAGONAL AND OFF-DIAGONAL REGULARIZATION TERMS. " ✓" INDICATES THE SELECTED REGULARIZATION TERM USED FOR OPTIMIZATION.

Method	Diagonal	Off-diagonal	Drone → Satellite		Satellite → Drone	
			R@1	AP	R@1	AP
Baseline+Ours			57.09	61.88	73.89	58.73
	✓		64.96	69.28	79.03	66.80
		✓	63.72	67.86	78.60	63.34
	✓	✓	**69.77**	**73.73**	**81.46**	**70.45**
LPN+Ours			75.93	79.14	86.45	74.49
	✓		78.58	81.49	86.59	78.22
		✓	78.09	80.94	85.73	76.85
	✓	✓	**81.51**	**84.11**	**88.30**	**79.38**
Swin-B+Ours			84.15	86.62	90.30	83.55
	✓		85.89	88.05	91.16	85.60
		✓	82.32	84.99	89.73	83.03
	✓	✓	**86.41**	**88.41**	**91.30**	**86.02**

TABLE VII
SENSITIVITY ANALYSIS FOR SETTING DIFFERENT WEIGHT INTERVALS. β_1 AND β_2 DENOTE COEFFICIENTS USED TO ADJUST INTERVALS OF ω_1 AND ω_2, RESPECTIVELY.

Method	β_1	β_2	Drone → Satellite		Satellite → Drone	
			R@1	AP	R@1	AP
Baseline+Ours	1	1	**69.77**	**73.73**	**81.46**	**70.45**
	2	1	66.10	70.26	77.60	67.02
	1	2	69.05	72.96	80.46	68.82
	2	2	67.45	71.61	79.17	67.63
LPN+Ours	1	1	**81.51**	**84.11**	**88.30**	**79.38**
	2	1	80.17	82.91	87.45	79.61
	1	2	80.86	83.53	88.59	80.46
	2	2	80.33	83.05	86.73	79.15
Swin-B+Ours	1	1	**86.41**	**88.41**	**91.30**	**86.02**
	2	1	85.11	87.29	89.02	84.49
	1	2	86.16	88.26	91.16	85.86
	2	2	85.85	87.98	90.01	85.28

regularization (DWDR). In DWDR, γ_1 and γ_2 are two key parameters that flexibly adjust the rate at which well-regressed elements of the Pearson cross-correlation coefficient matrix are down-weighted. When $\gamma_1 = 0$ and $\gamma_2 = 0$, DWDR does not apply two dynamic weights and can be viewed as Barlow Twins [43]. We empirically tune γ_1 and γ_2, and the related results are detailed in Table V. We first observe that applying one dynamic weight, *i.e.*, only $\gamma_1 = 1$ or only $\gamma_2 = 1$, achieves similar results to Barlow Twins. The limited performance improvement reflects that ignoring poorly-regressed diagonal and off-diagonal elements both induce the optimization plateau. When both γ_1 and γ_2 are set to 1, *i.e.*, using two dynamic weights, we obtain the best results. Specifically, compared with deploying Barlow Twins as regularization ("Baseline + *BT*"), our method ("Baseline + Ours") boosts R@1 from 67.91% to 69.77% (+1.86%) and AP from 71.99% to 73.73% (+1.74%) on Drone → Satellite, and goes up R@1 from 80.17% to 81.46% (+1.29%) and AP from 68.03% to 70.45% (+2.42%) on Satellite → Drone. When $\gamma_1 = 2$ and $\gamma_2 = 2$, the performance gains slightly degrades. A reasonable speculation is that large *focusing* parameters γ_1 and γ_2 cause the importance of poorly-regressed elements in the optimization process to be excessively reduced as well. To further verify the robustness of selected parameters, we conduct the same experiments in "LPN [4] + Ours" and "Swin-B [23] + Ours" and find the same conclusion. That is, when both γ_1 and γ_2 are set to 1, models achieve competitive results. Therefore, we choose $\gamma_1 = 1$ and $\gamma_2 = 1$ as default *focusing* parameters of DWDR. All three groups of experiments also support that DWDR is more effective than Barlow Twins for motivating networks to learn low-redundancy visual embeddings.

Effect of diagonal and off-diagonal regularization terms.

As shown in Eq. 8, DWDR consists of the diagonal and off-diagonal regularization terms. The diagonal regularization term boosts the positive linear correlation between two visual embeddings with the same geo-tag, and the off-diagonal regularization term tends to minimize redundancy among embedding channels. Table VI recapitulates the results by applying different regularization terms to three methods. First, applying either the diagonal or the off-diagonal regularization term alone yields competitive results compared to utilizing neither. We then observe that the results obtained by independently harnessing two regularization terms are close and much lower than the method of applying both regularization terms jointly. The experiments reveal two points. One is that the diagonal and off-diagonal regularization terms are equally important and can be deployed separately to improve model performance. The other point is that two regularization terms can complement each other and further facilitate the extraction of discriminative features when used together.

Effect of different weight intervals. The default interval of our dynamic weights is 0 to 1. We set $\beta_1\omega_1 \in [0, \beta_1]$ and $\beta_2\omega_2 \in [0, \beta_2]$ to study the sensitivity of utilizing different weight intervals. β_1 and β_2 are two coefficients that determine the maximum right boundary of the dynamic weight. We roughly select four sets of values and conduct experiments on three methods. Two observations can be found in Table VII. First, we achieve the best performance in the default condition (*i.e.*, $\beta_1 = 1$ and $\beta_2 = 1$). In addition, the model accuracy is almost unchanged when only $\beta_2 = 2$ but decreases slightly when $\beta_1 = 2$. The experimental results show that ω_1 is more sensitive to the weight interval. A main factor is the longer regression distance for diagonal elements during optimization. When the dynamic weight is changed, the loss generated by the diagonal regularization term fluctuates more and destroys the optimization balance of the training.

Effect of our sampling strategy and DWDR. Our symmetric sampling strategy is a combination of the drone-view based sampling and the satellite-view based sampling. To discuss the effectiveness of our sampling strategy, we conduct three groups of experiments under the condition of only changing the sampling strategy. Meanwhile, in each group of experiments, we study the effectiveness of DWDR. The experimental results are shown in Table VIII. We observe first that utilizing DWDR alone does not give comparable results to the baseline ("Instance Loss [1]") shown in Table II. However, when applied in conjunction with Instance Loss, DWDR significantly improves the performance of the network, regardless of the sampling strategy. Experiments within each group verify from the side that DWDR concentrates more on the redundant channels of the embedding itself rather than the distance between cross-view embeddings. Second, when only Instance Loss is harnessed, the drone-view based and the satellite-view based sampling acquire similar results to the baseline ("Instance Loss") applying random sampling. In contrast, the symmetric sampling strategy obtains the best geo-localization accuracy. Furthermore, the symmetric sampling strategy is also the most competitive in the other two experimental settings, *i.e.*, DWDR alone and Instance Loss plus DWDR. The significant performance increment demonstrates that the symmetric sampling strategy as a by-product is effective.

TABLE VIII
EFFECT OF THE SYMMETRIC SAMPLING STRATEGY AND THE DYNAMIC WEIGHTED DECORRELATION REGULARIZATION (DWDR).

Method	Instance Loss	DWDR	Drone → Satellite		Satellite → Drone	
			R@1	AP	R@1	AP
Drone-view based sampling	✓		60.86	65.56	73.89	59.05
		✓	22.61	27.84	35.38	20.98
	✓	✓	**66.08**	**70.22**	**77.60**	**65.48**
Satellite-view based sampling	✓		58.54	63.10	73.61	58.49
		✓	24.25	29.37	39.09	22.83
	✓	✓	**65.06**	**69.16**	**76.46**	**65.27**
The symmetric sampling strategy (**Ours**)	✓		64.74	68.96	77.75	64.32
		✓	34.38	40.02	50.07	34.28
	✓	✓	**69.77**	**73.73**	**81.46**	**70.45**

Effect of the dimension of visual embeddings. We deploy the final visual embeddings with different dimensions in geo-localization to investigate the effect of embedding dimensions on retrieval accuracy. The experimental results of the baseline and "Baseline + Ours" are shown in Table IX. We observe that with the increment of the dimension, both the baseline [1] and "Baseline + Ours" have a persistent improvement since the visual embedding possesses more information capacity. However, the performance of the two methods encounters the bottleneck when the feature dimension is 512. As the dimension of the feature increases to 1024, the performance of the baseline decreases significantly, and the performance of "Baseline + Ours" tends to stabilize. The experimental results reflect two aspects from the side. One is that features with too high dimensions are prone to redundant channels, which compromise the geo-localization accuracy of models. The other is that our method can encourage networks to learn low-redundancy embeddings and improve the robustness of the model. In addition, as shown in Figure 4, we notice that when the dimension raises from 64 to 128, the baseline achieves a higher growth rate than "Baseline + Ours". The short-dimensional features with small information capacity limit the performance of models. We speculate that our method allows the model to include more primary discriminative patterns in the limited feature dimension to mitigate the negative effects of insufficient information capacity. Therefore, when the feature dimension increases, our method produces fewer performance fluctuations.

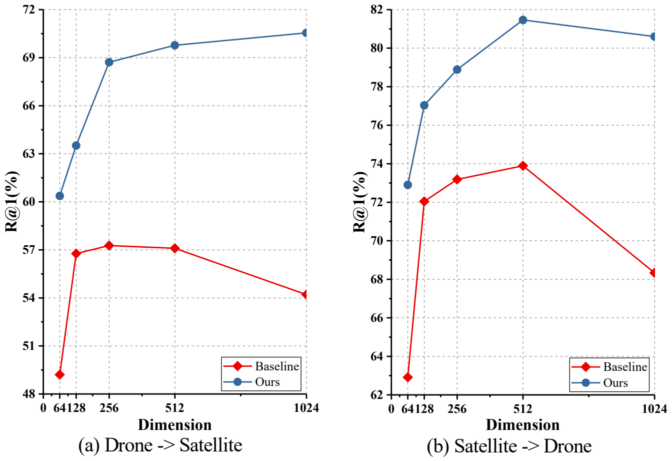
Effect of DWDR under different loss functions. Our baseline applies the instance loss [36], [37] to optimize the network while other loss functions are available. The triplet loss and the soft margin triplet loss are broadly utilized in previous works [34], [17], [18], [33]. We also evaluate our DWDR by deploying baselines adopting these two loss functions. The margin value of the triplet loss is 0.3, and experimental results are shown in Table X. We notice that both baselines combined with DWDR gain improved retrieval accuracy on the "Drone → Satellite" task and the "Satellite → Drone" task of University-1652.

Effect of the intra-view DWDR. Our method applies the

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2024.3491757

TABLE IX
ABLATION STUDY OF CROSS-VIEW GEO-LOCALIZATION APPLYING VISUAL FEATURES WITH DIFFERENT DIMENSIONS. "DIM" DENOTES THE DIMENSION OF FEATURES.

Method	Dim	Drone → Satellite		Satellite → Drone	
		R@1	AP	R@1	AP
Baseline [1] (Instance Loss)	64	49.20	54.36	62.91	49.68
	128	56.76	61.74	72.04	58.14
	256	**57.26**	**62.17**	73.18	58.70
	512	57.09	61.88	**73.89**	**58.73**
	1024	54.20	59.20	68.33	55.37
Baseline [1]+Ours	64	60.37	65.03	72.90	60.31
	128	63.51	68.05	77.03	64.57
	256	68.71	72.72	78.89	68.44
	512	69.77	73.73	**81.46**	70.45
	1024	**70.55**	**74.56**	80.60	**70.51**



(a) Drone -> Satellite

(b) Satellite -> Drone

Fig. 4. Impact of the dimension of features. The R@1 accuracy between the baseline and our method is compared. The red line refers to the baseline [1], and our method is shown using the blue line. (a) The drone-view target localization task (Drone → Satellite). (b) The drone navigation task (Satellite → Drone). When the feature dimension changes from 128 to 64, the performance of our method drops less than the baseline.

cross-view DWDR, in which the Pearson cross-correlation coefficient matrix is computed employing cross-view images. The intra-view DWDR means that the Pearson cross-correlation coefficient matrix of DWDR is calculated employing two distorted images from the same platform generated by different data augmentations. In experiments, the method only utilizing the symmetric sampling strategy is treated as the baseline, and the comparison results are shown in Table XI. We observe that the baseline combined with the intra-view DWDR gains a slight increment. Although the intra-view DWDR also encourages the network to learn independent embedding channels, our cross-view DWDR significantly outperforms applying the intra-view DWDR. It is because the cross-view DWDR is aligned with the cross-view retrieval test setting, which considers embeddings from different platforms for the geo-localization task. It also explains the limited performance increase of applying both intra-view and cross-view DWDR, which relies on the cross-view DWDR.

TABLE X
ABLATION STUDY OF DWDR UNDER DIFFERENT LOSS FUNCTIONS. "M" DENOTES THE MARGIN OF THE TRIPLET LOSS.

Method	University-1652			
	Drone → Satellite		Satellite → Drone	
	R@1	AP	R@1	AP
Triplet Loss ($M = 0.3$) [49]	52.16	57.47	63.91	52.24
Soft Margin Triplet Loss [34]	53.67	58.69	67.90	54.76
Triplet Loss ($M = 0.3$) + DWDR	54.14	59.28	67.90	54.76
Soft Margin Triplet Loss + DWDR	57.75	62.58	69.33	57.46

TABLE XI
ABLATION STUDY OF THE SYMMETRIC SAMPLING STRATEGY COMBINED WITH DIFFERENT DWDR.

Method	University-1652			
	Drone → Satellite		Satellite → Drone	
	R@1	AP	R@1	AP
Symmetric sampling **(Baseline)**	64.74	68.96	77.75	64.32
+ Intra-view DWDR	65.31	69.57	79.17	65.74
+ Cross-view DWDR	69.77	73.73	81.46	70.45
+ Intra & Cross-view DWDR	69.81	73.68	82.45	70.86

E. Qualitative Results

We visualize some heatmaps generated by the baseline and our method as an extra qualitative evaluation. Figure 5 shows the acquired heatmaps in the drone and satellite platforms of University-1652. Images in University-1652 possess an obvious geographic target. Compared with the baseline [1], our method activates a wider range of geographic target regions. In addition, we show some retrieval results on different datasets (see Figure 6). University-1652 supports two tasks. In the drone-view target localization task, the drone-platform image is the query, and in the drone navigation task, the satellite-platform image is the query. The retrieval results of two tasks are shown in Figure 6 (I) and (II). Figure 6 (III) and (IV) show the retrieval results of the ground-to-satellite localization task on CVUSA and CVACT. Given a randomly selected test query, we notice that the proposed method has successfully retrieved the most relevant results from the candidate gallery.

V. CONCLUSION

In this paper, we propose a dynamic weighted decorrelation regularization (DWDR) to achieve the cross-view geo-

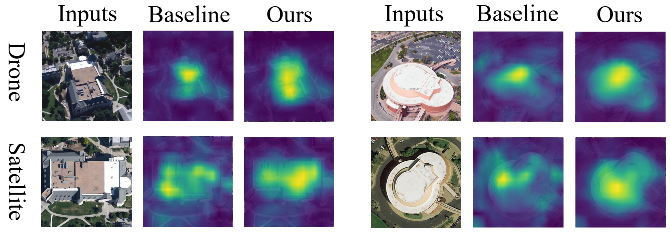


Fig. 5. Visualization of heatmaps. Heatmaps are produced by the baseline [1] and ours on different platforms of University-1652, *i.e.*, the drone platform and the satellite platform.

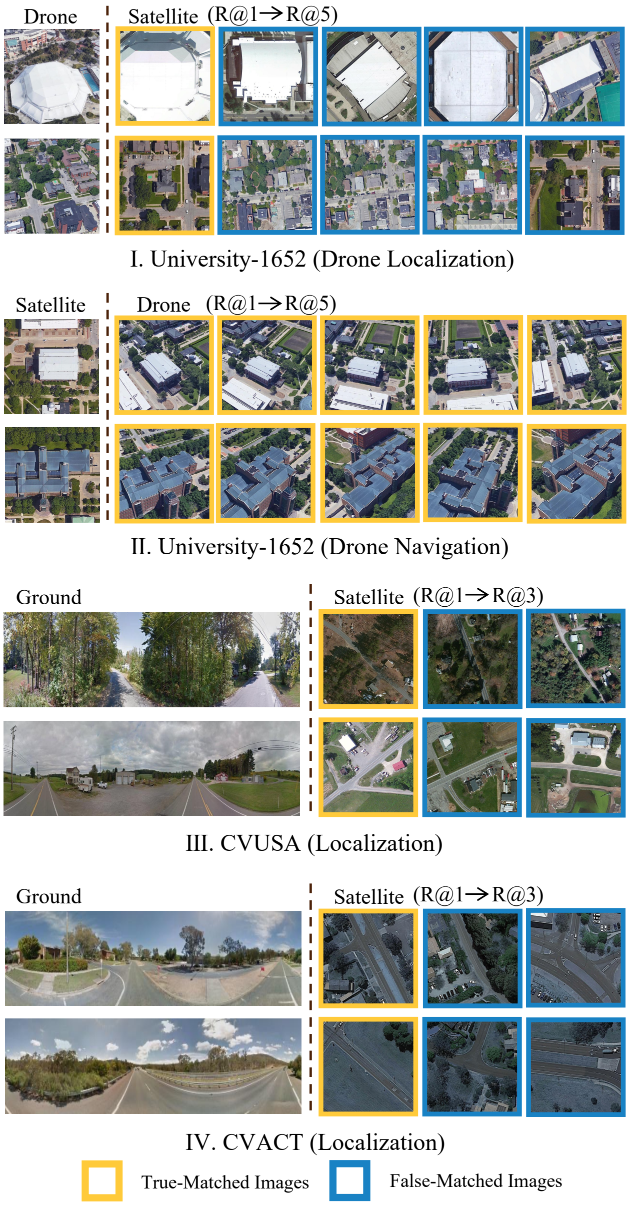


Fig. 6. Qualitative image retrieval results in different datasets. (I) and (II) show Top-5 retrieval results on University-1652. Different query images indicate the different tasks. (I) is the drone-view target localization task, and (II) is the drone navigation task. (III) and (IV) exhibit Top-3 retrieval results of geographic localization on CVUSA and CVACT, respectively. The true matches are in yellow boxes, and the false matches are highlighted by blue boxes.

localization. DWDR reduces the redundancy of visual embeddings by motivating the network to learn independent embedding channels. Specifically, DWDR sets dynamic weights to focus on the poorly-regressed elements when constraining the objective matrix to be as close as possible to the identity matrix. As a by-product of DWDR, the cross-view symmetric sampling strategy is introduced to balance the example number from different platforms in a training batch. The extensive experiments on four datasets, *i.e.*, University-1652, CVUSA, CVACT and VIGOR, demonstrate that our method can learn discriminative embeddings, which significantly improve the retrieval accuracy. Moreover, our method also acquires competitive results with the extremely short feature.

References

[1] Z. Zheng, Y. Wei, and Y. Yang, "University-1652: A multi-view multi-source benchmark for drone-based geo-localization," in *ACM International Conference on Multimedia*, 2020, doi: 10.1145/3394171.3413896.

[2] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Neural Information Processing Systems*, 2016.

[3] X. Ding, X. Zhang, J. Han, and G. Ding, "Scaling up your kernels to 31x31: Revisiting large kernel design in cnns," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[4] T. Wang, Z. Zheng, C. Yan, J. Zhang, Y. Sun, B. Zheng, and Y. Yang, "Each part matters: Local patterns facilitate cross-view geo-localization," *IEEE Transactions on Circuits and Systems for Video Technology*, 2022, doi:10.1109/TCSVT.2021.3061265.

[5] Z. Zheng, Y. Shi, T. Wang, J. Liu, J. Fang, Y. Wei, and T.-s. Chua, "Uavm '23: 2023 workshop on uavs in multimedia: Capturing the world from a new perspective," in *ACM International Conference on Multimedia*, 2023.

[6] G. Liu, C. Li, S. Zhang, and Y. Yuan, "Vl-mfl: Uav visual localization based on multisource image feature learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–12, 2024.

[7] Z. Zheng, T. Ruan, Y. W. Wei, Y. Yang, and M. Tao, "Vehiclenet: Learning robust visual representation for vehicle re-identification," *IEEE Transactions on Multimedia (TMM)*, 2020, doi: 10.1109/TMM.2020.3014488.

[8] Q. Chen, T. Wang, Z. Yang, H. Li, R. Lu, Y. Sun, B. Zheng, and C. Yan, "Sdpl: Shifting-dense partition learning for uav-view geo-localization," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2024.

[9] F. Ge, Y. Zhang, Y. Liu, G. Wang, S. Coleman, D. Kerr, and L. Wang, "Multibranch joint representation learning based on information fusion strategy for cross-view geo-localization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–16, 2024.

[10] H. Lv, H. Zhu, R. Zhu, F. Wu, C. Wang, M. Cai, and K. Zhang, "Direction-guided multiscale feature fusion network for geo-localization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–13, 2024.

[11] X. Wang, Z. Zheng, Y. He, F. Yan, Z. Zeng, and Y. Yang, "Progressive local filter pruning for image retrieval acceleration," *IEEE Transactions on Multimedia*, vol. 25, pp. 9597–9607, 2023.

[12] F. Ge, Y. Zhang, L. Wang, W. Liu, Y. Liu, S. Coleman, and D. Kerr, "Multilevel feedback joint representation learning network based on adaptive area elimination for cross-view geo-localization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–15, 2024.

[13] S. Li, M. Hu, X. Xiao, and Z. Tu, "Patch similarity self-knowledge distillation for cross-view geo-localization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 6, pp. 5091–5103, 2024.

[14] Q. Wu, Y. Wan, Z. Zheng, Y. Zhang, G. Wang, and Z. Zhao, "Camp: A cross-view geo-localization method using contrastive attributes mining and position-aware partitioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–14, 2024.

[15] J. Lin, Z. Zheng, Z. Zhong, Z. Luo, S. Li, Y. Yang, and N. Sebe, "Joint representation learning and keypoint detection for cross-view geo-localization," *IEEE Transactions on Image Processing*, 2022.

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

[17] L. Liu and H. Li, "Lending orientation to neural networks for cross-view geo-localization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[18] Y. Shi, X. Yu, L. Liu, T. Zhang, and H. Li, "Optimal feature transport for cross-view image geo-localization," in *AAAI Conference on Artificial Intelligence*, 2020.

[19] W.-J. Ahn, S.-Y. Park, D.-S. Pae, H.-D. Choi, and M.-T. Lim, "Bridging viewpoints in cross-view geo-localization with siamese vision transformer," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–12, 2024.

[20] H. B. Barlow *et al.*, "Possible principles underlying the transformation of sensory messages," *Sensory communication*, vol. 1, no. 01, 1961.

[21] L. Ding, J. Zhou, L. Meng, and Z. Long, "A practical cross-view image matching method between uav and satellite for uav-based geo-localization," *Remote Sensing*, vol. 13, no. 1, p. 47, 2021.

[22] M. Dai, J. Hu, J. Zhuang, and E. Zheng, "A transformer-based feature segmentation and region alignment method for uav-view geo-localization," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[23] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *IEEE International Conference on Computer Vision*, 2021.

[24] M. Bansal, H. S. Sawhney, H. Cheng, and K. Daniilidis, "Geo-localization of street views with aerial image databases," in *ACM International Conference on Multimedia*, 2011.

[25] T.-Y. Lin, S. Belongie, and J. Hays, "Cross-view image geolocalization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[26] F. Castaldo, A. R. Zamir, R. Angst, F. A. N. Palmieri, and S. Savarese, "Semantic cross-view matching," in *IEEE International Conference on Computer Vision Workshops*, 2015.

[27] S. Workman and N. Jacobs, "On the location dependence of convolutional neural network features," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[29] S. Workman, R. Souvenir, and N. Jacobs, "Wide-area image geolocalization with aerial reference imagery," in *IEEE International Conference on Computer Vision*, 2015.

[30] T.-Y. Lin, Y. Cui, S. Belongie, and J. Hays, "Learning deep representations for ground-to-aerial geolocalization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[31] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[32] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[33] N. N. Vo and J. Hays, "Localizing and orienting street views using overhead imagery," in *European Conference on Computer Vision*, 2016.

[34] S. Hu, M. Feng, R. M. Nguyen, and G. Hee Lee, "Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[35] S. Cai, Y. Guo, S. Khan, J. Hu, and G. Wen, "Ground-to-aerial image geo-localization with a hard exemplar reweighting triplet loss," in *IEEE International Conference on Computer Vision*, 2019.

[36] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, and Y.-D. Shen, "Dual-path convolutional image-text embeddings with instance loss," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, no. 2, pp. 1–23, 2020, doi: 10.1145/3383184.

[37] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *IEEE International Conference on Computer Vision*, 2017, doi: 10.1109/ICCV.2017.405.

[38] 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 Recognition*, vol. 152, p. 110363, 2024.

[39] Y. Shi, X. Yu, D. Campbell, and H. Li, "Where am i looking at? joint location and orientation estimation by cross-view matching," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[40] Y. Bengio and J. Bergstra, "Slow, decorrelated features for pretraining complex cell-like networks," in *Neural Information Processing Systems*, 2009.

[41] M. Cogswell, F. Ahmed, R. Girshick, L. Zitnick, and D. Batra, "Reducing overfitting in deep networks by decorrelating representations," in *International Conference on Learning Representations*, 2016.

[42] Y. Sun, L. Zheng, W. Deng, and S. Wang, "Svdnet for pedestrian retrieval," in *IEEE International Conference on Computer Vision*, 2017.

[43] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *International Conference on Machine Learning*, 2021.

[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[46] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[47] M. Zhai, Z. Bessinger, S. Workman, and N. Jacobs, "Predicting ground-level scene layout from aerial imagery," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[48] S. Zhu, T. Yang, and C. Chen, "Vigor: Cross-view image geo-localization beyond one-to-one retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[49] G. Chechik, V. Sharma, U. Shalit, and S. Bengio, "Large scale online learning of image similarity through ranking," in *Iberian Conference on Pattern Recognition and Image Analysis*, 2009.

[50] X. Tian, J. Shao, D. Ouyang, and H. T. Shen, "Uav-satellite view synthesis for cross-view geo-localization," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[51] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[52] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Neural Information Processing Systems*, 2019.

[53] K. Regmi and M. Shah, "Bridging the domain gap for ground-to-aerial image matching," in *IEEE International Conference on Computer Vision*, 2019.

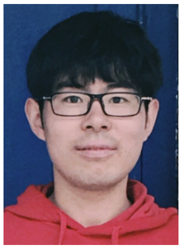
[54] Y. Shi, X. Yu, D. Campbell, and H. Li, "Where am i looking at? joint location and orientation estimation by cross-view matching," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[55] J. Li, C. Yang, B. Qi, M. Zhu, and N. Wu, "4scig: A four-branch framework to reduce the interference of sky area in cross-view image geo-localization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–18, 2024.

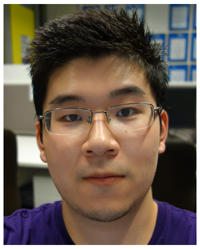
[56] S. Zhu, M. Shah, and C. Chen, "Transgeo: Transformer is all you need for cross-view image geo-localization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[57] J. Kwon, J. Kim, H. Park, and I. K. Choi, "Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks," in *International Conference on Machine Learning*, 2021.

[58] S. Zhu, T. Yang, and C. Chen, "Revisiting street-to-aerial view image geo-localization and orientation estimation," in *IEEE Winter Conference on Applications of Computer Vision*, 2021.



Tingyu Wang is a research assistant at School of Communication Engineering, Hangzhou Dianzi University, Hangzhou, China. He received his Ph.D. degree from the Lab of Intelligent Information Processing, Hangzhou Dianzi University, in 2023. His research interests include deep learning, image retrieval and remote sensing.



Zhedong Zheng is an assistant professor with the University of Macau. He was a research fellow at School of Computing, National University of Singapore. He received the Ph.D. degree from the University of Technology Sydney, Australia, in 2021 and the B.S. degree from Fudan University, China, in 2016. He received the IEEE Circuits and Systems Society Outstanding Young Author Award of 2021. He has organized a special session on reliable retrieval at ICME'22, two workshops at ACM MM'23 and one workshop at ACM ICMR'24. Besides, he is invited as a keynote speaker at CVPR'20, CVPR'21, a tutorial speaker at ACM MM'22. He also serves as an area chair at ACM MM'24.

Zunjie Zhu received the B.S. degree in electronic and information engineering and the Ph.D. degree in automation from Hangzhou Dianzi University, Hangzhou, China, in 2016 and 2022, respectively. He is currently an Assistant Professor with the School of Communication Engineering, Hangzhou Dianzi University. His research interests include 3-D vision, simultaneous localization and mapping (SLAM), and image restoration.

Yaoqi Sun received the B.S. degree from Zhejiang University of Science and Technology, Zhejiang, China, in 2012. He is currently pursuing a Ph.D. degree with Hangzhou Dianzi University, Zhejiang. He is an Assistant Research Fellow with School of Communication Engineering, Hangzhou Dianzi University. His research interests include intelligent information processing, machine learning, and pattern recognition.

Chenggang Yan received the B.S. degree in control science and engineering from Shandong University, Shandong, China, in 2008, and the Ph.D. degree in computer science from Chinese Academy of Sciences University, Beijing, China, in 2013. He is currently a Professor at School of Communication Engineering, Hangzhou Dianzi University. His research interests include computational photography, pattern recognition and intelligent system.

Yi Yang received the Ph.D. degree in computer science from Zhejiang University,Hangzhou, China, in 2010. He is currently a professor with Zhejiang University,Zhejiang, China. He was a Post-Doctoral Research with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA. His current research interest includes machine learning and its applications to multimedia content analysis and computer vision, such as multimedia indexing and retrieval, video analysis and video semantics understanding.