Geo-Localization via Ground-to-Satellite Cross-View Image Retrieval

Zelong Zeng, Zheng Wang, *Member, IEEE,* Fan Yang, and Shin'ichi Satoh, *Member, IEEE*

***Abstract*—The large variation of viewpoint and irrelevant content around the target always hinder accurate image retrieval and its subsequent tasks. In this paper, we investigate an extremely challenging task: given a ground-view image of a landmark, we aim to achieve cross-view geo-localization by searching out its corresponding satellite-view images. Specifically, the challenge comes from the gap between ground-view and satellite-view, which includes not only large viewpoint changes (some parts of the landmark may be invisible from front view to top view) but also highly irrelevant background (the target landmark tend to be hidden in other surrounding buildings), making it difficult to learn a common representation or a suitable mapping.**

To address this issue, we take advantage of drone-view information as a bridge between ground-view and satellite-view domains. We propose a Peer Learning and Cross Diffusion (PLCD) framework. PLCD consists of three parts: 1) a peer learning across ground-view and drone-view to find visible parts to benefit ground-drone cross-view representation learning; 2) a patch-based network for satellite-drone cross-view representation learning; 3) a cross diffusion between ground-drone space and satellite-drone space. Extensive experiments conducted on the University-Earth and University-Google datasets show that our method outperforms state-of-the-arts significantly.

I. Introduction

CROSS-VIEW image retrieval consist in retrieving the most relevant images from different platforms, which has received significant attention in recent years due to a large number of potential applications [1], [2], [3], [4], [5], [6]. [1] proposed a sketch-based image retrieval method, [2] focused on image-text cross-modal retrieval, [3] analysed cross-modal fashion retrieval tasks, [4] proposed a hashing approach for large-scale multimedia search tasks, and [5] used a multi-view hashing approach for social image retrieval tasks. Most previous works attempt to find a feature representation that is robust to the view variations. They have leveraged the deep neural networks together with metric learning strategies to learn discriminative representations for images [7], [8], [9], [10], [11], [12], [13], [14], [15], [16]. They designed different network architectures (*e.g.*, two-streams architectures [9], [10], [11], [12]) and aggregated layers (*e.g.*, NetVLAD [8] and GeM [13]) for geo-localization task. [15], [16] used manifold diffusion for re-ranking and enhanced the evaluation of image retrieval. Most of them also involved the spatial alignment [17], [18], [19] and the attention mechanism [12] in designing networks.

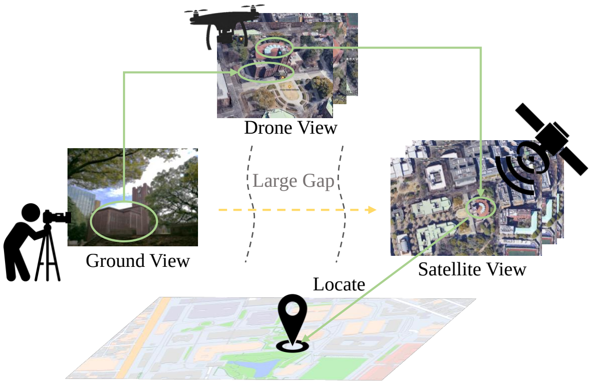


Fig. 1. An illustration of the task: given a ground-view image of a landmark, we aim to achieve cross-view geo-localization by searching out its corresponding satellite-view images. There is a very big gap between ground-view and satellite-view images due to the invisibility of parts of landmark and highly irrelevant background. We introduce the drone-view images to fill this gap and achieve the ground-drone-satellite image retrieval.

In practice, ground-to-satellite image retrieval is beneficial to image-based geo-localization. For instance, given a ground-view image showing the target landmark, the system retrieves and locates the same landmark among candidate satellite-view images. It can facilitate the positioning devices, like GPS, to provide a more accurate localization result. Previous works of ground-to-satellite image retrieval tried to learn a common representation across different views. Besides, two datasets CVUSA [20] and CVACT [19] are widely used for evaluation while the ground-view images provided by them are both panorama images. We argue that 1) the viewpoint varies greatly between ground-view and satellite-view images, where most components of the landmark shown in the ground-view image are invisible to the satellite-view; 2) it is not practical for users to obtain panorama ground-view images, thus the afore-mentioned datasets are not suitable for the real-world setting.

To tackle these issues, we propose to utilize drone-view images to bridge the domain gap between the ground-view and the satellite-view. As Figure 1 shows, a drone-view image may share some common components of the landmark with a ground-view image as well as a satellite-view image. This makes it reasonable to learn representations for ground-

This work was supported by National Key R&D Program of China (No.2021YFC3320301) and by National Natural Science Foundation of China (No.62171325). *(Corresponding author: Zheng Wang)*

Z. Wang is with the School of Computer Science, National Engineering Research Center for Multimedia Software, Wuhan University, China. (e-mail: wangzwhu@whu.edu.cn)

Z. Zeng, F. Yang, and S. Satoh are with the Department of Information and Communication Engineering, Graduate School of Information Science and Technology, The University of Tokyo, Japan, and also with the Digital Content and Media Sciences Research Division, National Institute of Informatics, Japan. (e-mail: zzlbz@nii.ac.jp;yang@nii.ac.jp;satoh@nii.ac.jp)

Manuscript received XXX; revised XXX.

to-drone and drone-to-satellite respectively. Inspired by this observation, we design a Peer Learning and Cross Diffusion (PLCD) framework, which is an effective way to build a connection from ground-view domain to satellite-view domain. Specifically, we divide the process of ground-to-satellite image retrieval into three steps: 1) ground-to-drone (ground→drone) image retrieval, 2) drone-to-satellite (drone→satellite) image retrieval, 3) diffusion between ground-drone and satellite-drone spaces. In the first step, considering the label noise (for a ground image, the relative drone-view images do not necessarily share the same facet regions of the landmark, because drone-view images of the same landmark are captured from different drone-viewing directions) and the background noise (the target landmark tend to be hidden in other surrounding buildings), we design a peer learning network which allows ground-view domain and drone-view domain to assist each other, and find the region of the target to benefit learn ground-drone cross-view representation. In the second step, we design a patch-based network to learn satellite-drone representation. For the third step, we generate a cross-view relation graph by using a cross diffusion to connect ground-drone and satellite-drone spaces. In Section III, we introduce each part in more detail.

To verify the effectiveness of the proposed method, we conduct experiments on two datasets University-Earth and University-Google [10], PLCD achieves 40.87%/21.77% CMC@1% accuracies for the task of ground-to-drone retrieval which are higher than the baseline work [10] by 34.24%/16.57% on the University-Earth and University-Google datasets respectively. Our contributions can be summarized as follows:

- **A new strategy.** We raise a new strategy for ground-to-satellite image retrieval. The strategy is taking advantage of drone-view information as a bridge between ground-view and satellite-view domains to overcome the huge domain gap.
- **A novel method.** We propose a novel Peer Learning and Cross Diffusion (PLCD) framework, which builds mapping from ground-view domain to satellite-view domain. The method achieves great performance improvements on the University-Earth and University-Google datasets.

II. Related Work

A. Ground-to-Satellite Geo-Localization

Ground-to-satellite geo-localization (Ground-view image → Satellite-view image) has been attracting more attention in recent years due to a large number of potential applications. It's a task of image localization on a geo-referenced satellite map given a query ground-view image. Workman *et al.* [7] attempted to fine-tune a pre-trained CNN by reducing the feature distance between pairs of ground-view images and satellite-view images for the cross-view localization task. Zhai *et al.* [21] designed a modified Siamese Network by plugging the NetVLAD [8] layer, they demonstrated that siamese-structure and aggregation layer can make image descriptors robust against large viewpoint changes. To take one step further, Hu *et al.* [9] proposed a weighted soft margin ranking loss function that not only speeds up the training convergence but also improved the final matching accuracy. Shi *et al.* [14] proposed CVFT layer which can transport features from one domain to the other, leading to more meaningful feature similarity comparison. Furthermore, Shi *et al.* [18] improved the performance of ground-to-satellite geo-localization through spatial alignment and attention mechanism. For these existing works, the ground-view images they used are all panorama images, which means each panorama image contains the same area as the relative satellite-view image. Most of the existing methods [22] focus on how to directly establish the association between images from different views, *i.e.*, to directly map images from different views to the same space. Different from them, our proposed method uses an intermediate view (drone view) to reduce the impact of large gaps between different views. In addition, we propose to use cross-diffusion to better connect the embedding spaces of different views.

B. Drone-to-Satellite Geo-Localization

Drone-to-satellite geo-localization (Drone-view image → Satellite-view image) is a new task of cross-view geo-localization. Given one drone-view image or video, the task aims to find the most similar satellite-view image to localize the target landmark in the satellite view. Because of the demand for drone applications, for instance, multi-view coordination [23], drone-view coordination [24], Drone-to-satellite geo-localization has been attracting more attention.

Zheng *et al.* [10] proposed Drone-to-satellite geo-localization task and a new multi-view multi-source benchmark for drone-based geo-localization, named University-1652. They demonstrated that the University-1652 helps the model to learn the viewpoint-invariant features and also has good generalization ability in the real-world scenario. Hu *et al.* [17] proposed an orientation-based method to align the patterns and introduce a new branch to extract aligned partial feature[25]. Moreover, they provided a style alignment strategy to reduce the variance in image style and enhance the feature unification. Inspired by the human visual system, Wang *et al.* [26] proposed a model which explicitly takes contextual patterns into consideration and leverages the surrounding environment around the target landmark, yielding better performance.

Some methods align images with different views at the image level, but their proposed methods require human preprocessing of the image, such as rotating the image to the same direction [17] or cropping the image [26]. In addition, most methods are based on global-level feature. Considering the presence of noisy samples, our approach uses peer learning, which allows the model to be adaptively align during training to extract the best matching local features and obtain better results.

C. Cross-Modal Image Retrieval

Cross-modal image retrieval allows using other types of query [27], [28], [29]. Some of the common types are text-to-image retrieval and sketch-to-image retrieval.

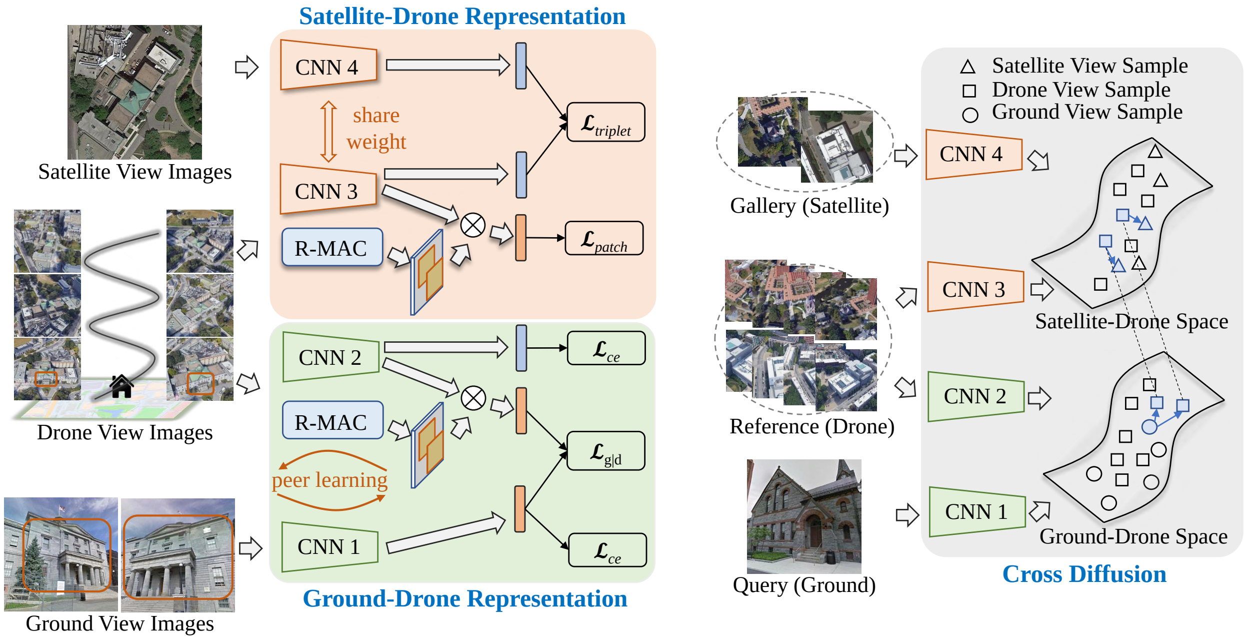


Fig. 2. The framework overview of our method. It consists of three parts, *i.e.*, 1) peer learning for the ground-drone cross-view representation; 2) patch-based model for the satellite-drone cross-view representation learning; 3) cross diffusion between ground-drone and satellite-drone spaces. We first learn the networks CNN_1 and CNN_2 for the ground-drone representation, as well as CNN_3 and CNN_4 for the satellite-drone representation. Then given a ground-view image of a landmark, cross diffusion is capable of achieving cross-view geo-localization by searching out corresponding satellite-view images, where we also use drone-view images (*no label*) acting as the reference set.

- **Text-to-image retrieval.** Most cross-model IR on text-to-image retrieval considers the case where the input query is formulated as an input image plus text describing the required modifications to the image. The main challenge of this task is to fuse the modified textual information into the source image and to evaluate the similarity between the fused image and the features of the desired image. Most approaches have focused on designing feature fusion layers and learning loss functions using different deep metrics. TIRG [30] used ResNet and RNN (LSTM) to extract features of the source image and modified text respectively, then fused them by using a gated residual connection, and finally used metric learning to learn the similarity metric. [31] extracted the features of the source image and that of the text by ResNet and RNN (GRU), and fused them through concatenation, then applied reinforcement learning to improve the quality of similarity metric learning.
- **Sketch-to-image retrieval.** The sketch-based image retrieval problem involves retrieving a hand-drawn sketch of a given object from a gallery of images of that class of objects. To bridge the gap between the sketch domain and the image domain, most approaches take the form of extracting the edge map of the image first, and then feeding the edge map and sketch into a deep network structure to extract features and perform matching. [32] used generative networks to enhance the discrimination of the extracted features. [33] applied the attention mechanism module and proposes a new loss function, namely HOLEF, to improve the performance.

In contrast to text/sketch-to-image retrieval, the data in the Ground-to-Satellite cross-view Image Retrieval (G2S IR) task are all images and no text/sketch data, so information fusion and information alignment for different modalities do not need to be considered in our work. Our work focuses more on feature mapping of multi-view images.

III. Our Method

In this section, we introduce the Peer Learning and Cross Diffusion (PLCD) framework (Figure 2). Due to the gap between ground-view and satellite-view, too large to cover by a common space, PLCD prepares “doable” common spaces for ground-drone and drone-satellite independently, and connect them via cross diffusion. PLCD framework can be decomposed into three main parts: 1) A peer learning across ground-view and drone-view to find visible correlative parts to enjoy benefit ground-drone local representation learning (Section III-A). The peer learning includes two steps, *i.e.,* train the senior peer and train the junior peer. 2) A patch-based model to obtain global representation across drone-view and satellite-view, and close the gap between ground-drone and drone-satellite spaces (Section III-B). 3) A cross diffusion to best utilize intermediate view, *i.e.*, drone view, for boosting ground-to-satellite retrieval as post processing (Section III-C).

Problem formulation. For ground-drone-satellite image retrieval, we have three subsets containing images of three different views, *i.e.*, ground-view, drone-view and satellite-view. We use $\mathcal{G} = \{\mathbf{g}\}$, $\mathcal{D} = \{\mathbf{d}\}$, and $\mathcal{S} = \{\mathbf{s}\}$ to denote the ground-view, drone-view, and satellite-view sets, respectively. For each image $\mathbf{g}$, $\mathbf{d}$, or $\mathbf{s}$, we use $y \in [1, \ldots, C]$

to represent its label, where C is the number of identities. Note that there are only identity labels and no annotated bounding box indicating the region of the landmark. The sets $\mathcal{G}$, $\mathcal{D}$ and $\mathcal{S}$ are divided into two parts, one part for training, the other for testing, and the training and testing parts have no overlapping identity. Given an image $\mathbf{g}$, our aim is to search out corresponding images that contain the same landmark in the set $\mathcal{S}$. In this paper, we learn cross-view representations via four CNN models. They are 1) the ground-drone representation model CNN$_1$ for ground-view images $f_{\theta_1}(\mathbf{g}) : \mathcal{G} \mapsto \mathbb{R}^d$; 2) the ground-drone representation model CNN$_2$ for drone-view images $f_{\theta_2}(\mathbf{d}) : \mathcal{D} \mapsto \mathbb{R}^d$; 3) the satellite-drone representation model CNN$_3$ for drone-view images $f_{\theta_3}(\mathbf{d}) : \mathcal{D} \mapsto \mathbb{R}^d$; and 4) the satellite-drone representation model CNN$_4$ for satellite-view images $f_{\theta_4}(\mathbf{s}) : \mathcal{G} \mapsto \mathbb{R}^d$, where d stands for the number of dimensions.

A. *Ground-Drone Cross-view Representation*

In major public benchmarks, each image is generally associated with a landmark identity, here called hard label. However, for each ground image, corresponding drone-view images do not necessarily share the same facet regions of the landmark. This is because drone-view images of the landmark are captured from different directions. If we train the model by directly using hard labels, the performance will not be so good. However, hard labels can help to discover potential valuable images that benefit training.

To make full use of valuable positive samples and focus on precise regions of targets, we propose a peering learning network which contains two branches, *i.e.*, the ground-view branch and the drone-view branch. Two branches do not share weights, but their final embedding features are pushed close to each other by constraints. Given a ground-view image, the *Detection* layer (RPN [34] pretrained on [35]) outputs the region of interest and the CNN$_1$ outputs the feature map. Next, a *Region of Interest* (ROI) pooling layer is applied to the detection result and outputs the feature. For the drone-view branch, the CNN$_2$ is used to extract activation features from drone-view images. Activation features are max-pooled in different regions of the image by using a multi-scale rigid grid with overlapping cells, called *R-MAC* [36] (as illustrated in Figure 3) layer. We use cross-entropy to optimize each branch respectively, and a consistency loss to learn the relation between two domains. The training process can be divided into two steps.

1) Step I: Train the Senior Peer: To avoid the effect from hard label noise, we first use the easiest samples. We feed 'easy' triplets as explained below to CNN$_1$ and CNN$_2$. Each triplet consists of one ground-view image $\mathbf{g}$, its easiest positive image $\mathbf{d}^*$ (share the same facet region of the landmark as $\mathbf{g}$) and its multiple difficult negative images $\{\mathbf{d}_j|_{j=1}^N\}$. Such triplets are sampled according to the image features f_{θ_1} or f_{θ_2}. The easiest positive image $\mathbf{d}^*$ is the top-1 ranking drone-view image from a positive image batch. To ensure that there exist easy positive images in each batch, we divide the drone-view images into D different sections according to drone's directions and randomly select one positive image from each section to form a batch. The most difficult negative images $\{\mathbf{d}_j|_{j=1}^N\}$ are the top-N ranking gallery image from a negative image batch. In this process, two branches assist each other and learn how to select the easiest positive and difficult negative samples. Here, we adopt a consistency loss [37] to project features from different domains to the same space. It is formulated as:

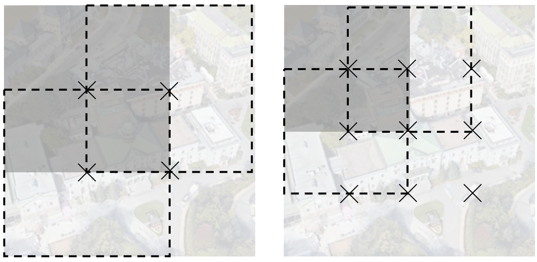


Fig. 3. An illustration of R-MAC layer. The figure shows example regions extracted at two different scales (l = 2, 3), the top-left region of each scale (gray colored region) and its neighboring regions (dashed borders). The centers of all regions are with a cross. At scale l, we sample $l \times l$ regions with the width $L/(l+1)$, where L indicates the length of square image.

$$\mathcal{L}_{\mathbf{g}|\mathbf{d}}(\theta_1,\theta_2)=\frac{\exp\left(-\|f_{\theta_1}(\mathbf{g})-f_{\theta_2}(\mathbf{d}^*)\|^2\right)}{\exp\left(-\|f_{\theta_1}(\mathbf{g})-f_{\theta_2}(\mathbf{d}^*)\|^2\right)+\sum_{j=1}^{N}\exp\left(-\|f_{\theta_1}(\mathbf{g})-f_{\theta_2}(\mathbf{d}_j)\|^2\right)}, \quad (1)$$

where $\|\cdot\|$ denotes the Euclidean Norm.

We also apply cross-entropy loss to improve the discriminative representation. The objective loss of step I $\mathcal{L}_{hard}$ combines the $\mathcal{L}_{\mathbf{g}|\mathbf{d}}$ and cross-entropy loss as:

$$\mathcal{L}_{hard} = \mathcal{L}_{\mathbf{g}|\mathbf{d}} - \sum_{j=1}^{C} p_j \log q_j, \quad (2)$$

where q_j is a score obtained by applying a softmax function to the logits of a classification layer for the sample, p_j is the ground truth for the sample, and C denotes the number of classes.

2) Step II: Train the Junior Peer: In step I, we only use the easiest positive samples and global information. To make full use of difficult positive samples and pay more attention to local information, we use the trained models of step I as a senior peer to guide a junior peer in step II. The structure of the junior peer is the same as the senior peer. Two peers both activate *R-MAC* layers.

We denote θ_1^{SP} and θ_2^{SP} are parameters of CNN$_1$ and CNN$_2$ in senior peer models, and we freeze them in step II. θ_1^{JP} and θ_2^{JP} are learnable parameters of CNN$_1$ and CNN$_2$ in junior peer models. Both of them are fed with doublets, and each doublet consists of one ground-view image $\mathbf{g}$, its multiple positive images $\{\mathbf{d}_i|_{i=1}^M\}$. Given a positive drone-view image $\mathbf{d}_i$, CNN$_2$ and *R-MAC* layer can output descriptors $\{f_{\theta_2}(\mathbf{d}_{i,1}), \cdots, f_{\theta_2}(\mathbf{d}_{i,m})\}$ corresponding to the m potential image sub-regions $\{\mathbf{d}_{i,1}, \cdots, \mathbf{d}_{i,m}\}$ proposed by *R-MAC* layer. The similarities between $\mathbf{g}$ and sub-regions are measured by dot products and normalized by a softmax function with temperature τ over the M positive images in the batch:

$$\begin{aligned}&\mathcal{S}\left(\mathbf{g},\mathbf{d}_1,\cdots,\mathbf{d}_M;\tau\right)=\operatorname{softmax}([\left\langle f_{\theta_1}(\mathbf{g}),f_{\theta_2}(\mathbf{d}_1)\right\rangle/\tau,\\&\left\langle f_{\theta_1}(\mathbf{g}),f_{\theta_2}(\mathbf{d}_{1,1})\right\rangle/\tau,\cdots,\left\langle f_{\theta_1}(\mathbf{g}),f_{\theta_2}(\mathbf{d}_{1,m})\right\rangle/\tau,\cdots,\\&\left\langle f_{\theta_1}(\mathbf{g}),f_{\theta_2}(\mathbf{d}_M)\right\rangle/\tau,\left\langle f_{\theta_1}(\mathbf{g}),f_{\theta_2}(\mathbf{d}_{M,1})\right\rangle/\tau,\cdots,\\&\left\langle f_{\theta_1}(\mathbf{g}),f_{\theta_2}(\mathbf{d}_{M,m})\right\rangle/\tau]),\end{aligned}\tag{3}$$

where $f_{\theta_1}(\mathbf{g})$ is the encoded feature representations of the ground-view image by the CNN$_1$ branch, $f_{\theta_2}(\mathbf{d}_i)$ and $f_{\theta_2}(\mathbf{d}_{i,j})$ are the encoded feature representations of the i-th gallery image and its j-th potential sub-region by the CNN$_2$ branch respectively. τ is temperature hyper-parameter less than 1, which makes the similarity vector sharper.

$\mathcal{S}^{SP}$ and $\mathcal{S}^{JP}$ are the similarities calculated by Equation 3 of senior peer model and junior peer model respectively. $\mathcal{S}^{SP}$ shows the distribution of the drone-view image and its sub-regions in senior peer model space, which serves as soft supervision to junior peer model via a cross-entropy loss:

$$\mathcal{L}_{soft}=-\sum_i\mathcal{S}_i^{SP}(\mathbf{g},\mathbf{d}_1,\cdots,\mathbf{d}_M;\tau)\log\left(\mathcal{S}_i^{JP}(\mathbf{g},\mathbf{d}_1,\cdots,\mathbf{d}_M;1)\right),\tag{4}$$

where $\mathcal{S}_i(\cdot)$ is the i-th element of softmax vector $\mathcal{S}(\cdot)$. At the same time, we also use hard loss $\mathcal{L}_{hard}$ (Equation 2) to supervise the junior peer model. $\mathcal{L}_{hard}$ and $\mathcal{L}_{soft}$ are jointly adopted in step II as:

$$\mathcal{L}_{\mathcal{G}-\mathcal{D}}=\mathcal{L}_{hard}+\lambda_1\mathcal{L}_{soft},\tag{5}$$

where the weights λ_1 is used for balancing these two losses.

B. Satellite-Drone Cross-view Representation

We design a patch-based network, which has two weight shared branches: CNN$_3$ and CNN$_4$. CNN$_3$ and CNN$_4$ can extract feature maps $\mathbf{m}_d$ and $\mathbf{m}_s$ from input drone-view image and satellite-view image respectively. Here, we use the pretrained peer learning model to supervise patch-based network for training, we found that this process can improve the performance of diffusion. Similar with we describe in Section III-A2, given a positive drone-view image $\mathbf{d}_i$, CNN$_2$ (from the pretrained peer learning model and we freeze it) and CNN$_3$ output the feature map $\mathbf{m}_{d_i}^{\theta_2}$ and $\mathbf{m}_{d_i}^{\theta_3}$ respectively. Then *R-MAC* layer proposes sub-regions from each feature map and transform each patch into features $\{f_{\theta_2}(\mathbf{d}_{i,1}),\cdots,f_{\theta_2}(\mathbf{d}_{i,m})\}$ and $\{f_{\theta_3}(\mathbf{d}_{i,1}),\cdots,f_{\theta_3}(\mathbf{d}_{i,m})\}$ respectively. This process can be written as:

$$\{f_s(\mathbf{d}_{i,j})|j\in\{1,\cdots,m\}\}=\mathcal{F}_{\text{R-MAC}}(\mathbf{m}_{d_i}^s),s\in\{\theta_2,\theta_3\},\tag{6}$$

where $\mathcal{F}_{\text{R-MAC}}$ indicates *R-MAC* layer, m is the number of sub-region proposed by *R-MAC* layer from each feature map.

Then we close the distance between each pair $\{f_{\theta_2}(\mathbf{d}_{i,j}),f_{\theta_3}(\mathbf{d}_{i,j})\}$ by Mean Squared Error (MSE) loss $\mathcal{L}_{mse}$:

$$\mathcal{L}_{patch}=\frac{1}{m}\sum_i\sum_{j=1}^{m}\mathcal{L}_{mse}(f_{\theta_2}(\mathbf{d}_{i,j}),f_{\theta_3}(\mathbf{d}_{i,j})),\tag{7}$$

where i indicates i-th drone-view image of the image batch.

Besides, we adopt a semi-hard triplet loss [38] $\mathcal{L}_{triplet}$ to project features from different domains to the same space. The objective function for this phase can be formulated as:

$$\mathcal{L}_{\mathcal{S}-\mathcal{D}}=\mathcal{L}_{triplet}+\lambda_2\mathcal{L}_{patch},\tag{8}$$

where the weights λ_2 is used for balancing these two losses.

C. Cross-diffusion

Once we obtained the feature representations, we need to connect the "doable" common spaces. Some prior works [16], [15], [39] have proven the effectiveness of diffusion when mining on manifolds. Here, we apply diffusion, called cross-diffusion, on heterogeneous manifolds (corss-domain) to retrieve the target in satellite-view images with a given ground-view image across the ground-drone and satellite-drone feature spaces.

As a graph-based random walk processing, the key of diffusion is the graph construction and the initialization. Specifically, the graph construction focuses on the local structure of feature space and defines the direction of random walk, while the initialization determines where to start the random walk. For the graph, we measure the similarities among image features $\{f_{\theta_3}(\mathbf{d})\}\cup\{f_{\theta_4}(\mathbf{s})\}$ in the satellite-drone feature space, which are used to form a transition matrix $\mathbf{S}\in\mathbb{R}^{(\#\mathcal{D}+\#\mathcal{S})\times(\#\mathcal{D}+\#\mathcal{S})}$ representing the graph. Essentially, the i-th column of the transition matrix records the probabilities that how likey the weight on the i-th node should diffuse to each of its neighbors, where the node stands for the image sample in our case. Now that we have obtained the graph for satellite-drone features, we conduct the initialization step by assigning weights to drone-view nodes who are nearest neighbors of the given ground-view query in the ground-drone feature space. The weights are stored in a state vector $\mathbf{f}\in\mathbb{R}^{\#\mathcal{D}+\#\mathcal{S}}$ whose initial state is defined as

$$\mathbf{f}_i^0=\begin{cases}(f_{\theta_1}(\mathbf{g}_q)^\top f_{\theta_2}(\mathbf{d}_i))^\gamma & i\in\text{NN}^{\text{ID}}(f_{\theta_1}(\mathbf{g}_q))\\0 & \text{otherwise}\end{cases},\tag{9}$$

where $\mathbf{g}_q$ stands for the ground-view query, $\text{NN}^{\text{ID}}(\cdot)$ refers to the indexes of nearest neighbors, and γ is a constant variable which is set to 3 by convention. Given the initial state vector and the trainsition matrix, the random walk iterates the following step:

$$\mathbf{f}^{t+1}=\alpha\mathbf{S}\mathbf{f}^t+(1-\alpha)\mathbf{f}^0,\quad\alpha\in(0,1),\tag{10}$$

which mathmatically converges to

$$\mathbf{f}^\infty\propto(\mathbf{I}-\alpha\mathbf{S})^{-1}\mathbf{f}^0\tag{11}$$

After convergence, we take the weights in $\mathbf{f}^\infty$ which correspond to the satellite-view images as clues for ranking.

D. Querying Process

The querying process can be divided in to two steps (as shown in Figure 4):

Feature Extraction. We use CNN$_1^{JP}$ to extract the features $f_{\theta_1}(\mathbf{g_q})$ of street-view query image $\mathbf{g_q}$, CNN$_2^{JP}$ and CNN$_3$

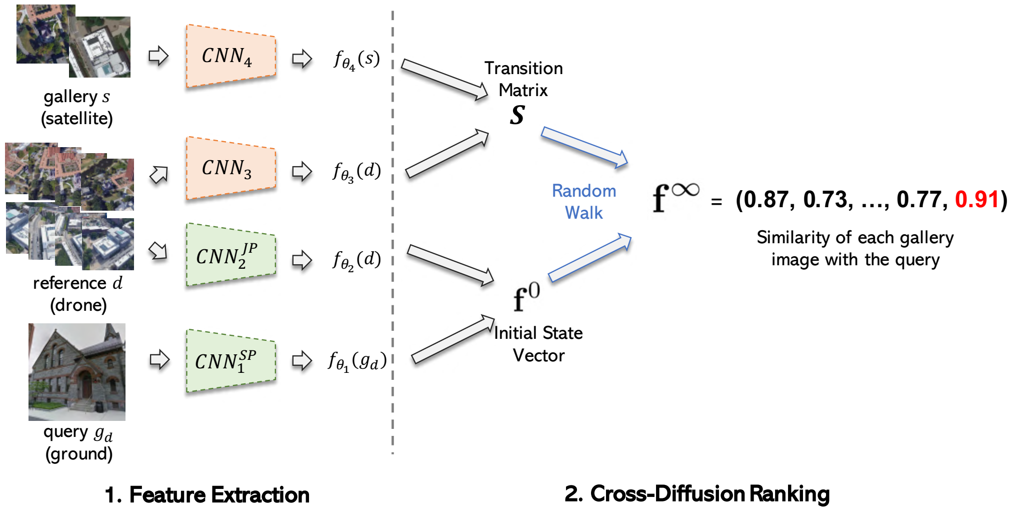


Fig. 4. Diagram to show the flow of querying process. It consists of two steps: 1) feature extraction and 2) cross-diffusion ranking. The output $\mathbf{f}^\infty$ corresponds to the satellite-view images as clues for ranking. The higher the weight means that the corresponding satellite-view gallery image is more similar to the drone-view query image.

to extract the features $f_{\theta_2}(\mathbf{d_i})$ and $f_{\theta_3}(\mathbf{d_i})$ of each drone-view reference image $\mathbf{d_i}$, and then CNN$_4$ to extract the features $f_{\theta_4}(\mathbf{s_j})$ of each satellite-view gallery image $\mathbf{s_j}$.

Cross-diffusion Ranking. We adopt diffusion to connect the "doable" common spaces (street-drone and drone-satellite). We first use features $f_{\theta_3}(\mathbf{d})$ and $f_{\theta_4}(\mathbf{s})$ to form the transition matrix $\mathbf{S}$. Next, we initialize the state vector $\mathbf{f^0}$. Finally, given the initial state vector $\mathbf{f^0}$ and the transition matrix $\mathbf{S}$, we adopt random walk to get the weight $\mathbf{f}^\infty$ which correspond to the satellite-view images as clues for ranking. The higher the weight means that the corresponding satellite-view gallery image is more similar to the drone-view query image.

IV. Experiments

A. Dataset and Evaluation Metric

Dataset. University-1652 [10] is a recently released multi-view multi-source dataset containing ground-view, drone-view and satellite-view data. It covers 1652 architectures of 72 universities around the world. The training set includes 701 buildings of 33 universities, and the test set includes the other 951 buildings of the rest 39 universities. There are no overlapping universities in the training and test set. Different from other similar multi-view datasets [20], [19], the ground-view images of University-1652 are not panorama ones. For each landmark, there are one satellite image, 54 drone-view images from different viewpoints and altitudes and several ground-view images from two different sets (one collected by google earth engine and another by google image search engine). Note that [10] has discussed the limitation of **University-1652**:

- To collect images of landmarks, they first obtained meta-data about university buildings from Wikipedia (*e.g.*, building name and university affiliation) and then encoded the building names into accurate geographical locations via Google Maps, *i.e.*, latitude and longitude. When we search for locations using Google Maps [40], Google Maps cannot always output precise results. Hence, they filter out buildings whose search results are unclear.
- Street-view images are collected from Google Maps (Google Street) and the image search engine (Google Image). When images are collected from the image search engine, the collection contains a large number of noisy results, including indoor environments and duplicate images. They applied a model trained on the Place dataset to detect the noisy results and filter them out. Due to accessibility, some buildings do not have good street-view photos, meaning that most street view images are captured by cameras on top of cars. As a result, there are occluded and irrelevant background street view images. Therefore, we applied the RPN [34] detection layer to extract the target regions in the proposed framework.

We divided the University-1652 dataset into two separate datasets: the first one, called **University-Earth**, consists of ground images collected by google earth, drone images collected at low altitude and and satellite images; the second one, called **University-Google**, consists of ground images collected by google image search engine, drone images collected at high altitude and satellite images. Since ground images and drone images are collected in different ways and from different altitudes, they can be considered as collected from different domains, and we usually use Google satellite image for positioning, so both datasets share the same satellite set. We believe that this setting is realistic.

Evaluation Metric. To indicate the performance, the standard Cumulative Matching Characteristics@K (CMC@K) values is adopted. CMC@K represents whether the correctly matched images in the top-K of the ranking list. A higher CMC@K score shows a better performance of the network. We also use mean Average Precision (mAP), which reflects the precision and recall rate of the retrieval performance. Especially, we focus on CMC@K performance on ground-drone image retrieval

TABLE I
THE COMPARISON WITH STATE-OF-THE-ARTS IN THE SETTING OF GROUND → SATELLITE RETRIEVAL. THE CMC@1, CMC@5, CMC@10, CMC@1% AND MAP VALUES (%) ARE REPORTED.

Method	**University-Earth**					**University-Google**				
	CMC@1	**CMC@5**	**CMC@10**	**CMC@1%**	**mAP**	**CMC@1**	**CMC@5**	**CMC@10**	**CMC@1%**	**mAP**
DELF [36] *w/o* $\mathcal{D}$	0.01	0.39	0.66	0.74	0.60	0.01	0.19	0.37	0.37	0.31
DELF [36]	0.12	0.50	0.93	0.93	0.87	0.05	0.26	0.53	0.55	0.53
R-MAC [41] *w/o* $\mathcal{D}$	1.09	3.61	6.59	6.94	2.19	0.74	2.79	5.23	5.54	1.66
R-MAC [41]	1.09	3.84	6.67	7.06	2.22	0.78	3.02	5.62	6.17	1.78
Str-CNN [10] *w/o* $\mathcal{D}$	0.74	2.79	4.85	5.66	1.70	0.27	1.90	3.76	4.07	1.09
Str-CNN [10]	1.01	3.22	6.01	6.63	2.08	0.54	2.40	4.77	5.20	1.47
Str-CNN [10] + Multi-loss	1.51	5.39	9.77	10.55	3.12	1.09	4.11	6.51	7.13	2.28
CVM-Net [9] *w/o* $\mathcal{D}$	0.35	1.05	2.09	2.29	0.88	1.24	3.61	6.01	6.24	2.26
CVM-Net [9]	1.78	4.69	8.61	9.42	3.18	0.70	4.15	6.36	6.79	1.89
Siam-FCANet50 [12] *w/o* $\mathcal{D}$	0.39	2.02	3.84	4.23	1.34	0.27	1.67	2.91	3.14	0.97
Siam-FCANet50 [12]	1.20	4.07	7.25	7.68	2.46	1.01	3.53	6.05	6.63	2.15
LPN [26] *w/o* $\mathcal{D}$	0.16	0.78	1.82	2.06	0.65	0.19	0.93	1.67	1.78	0.69
LPN [26]	0.74	3.10	4.69	5.04	1.70	0.62	2.60	4.54	5.04	1.57
Instance Loss [42] *w/o* $\mathcal{D}$	0.62	-	5.51	-	1.60	-	-	-	-	-
Instance Loss [42]	1.20	-	7.56	-	2.52	-	-	-	-	-
PLCD (Ours)	**9.15**	**27.66**	**38.83**	**40.87**	**14.16**	**4.26**	**10.62**	**20.40**	**21.77**	**7.63**

task, because of existing hard label noise in drone-view images and we only expect to find out those drone-view images share the same facet as the query image.

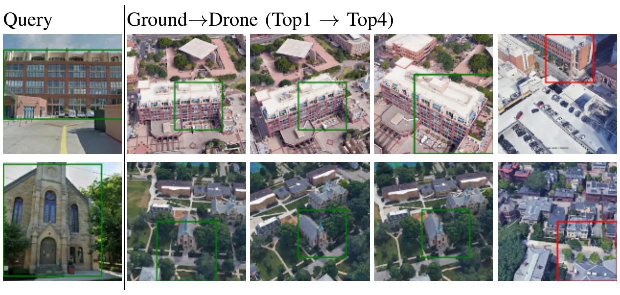


Fig. 5. The visualization of ground-to-drone image retrieval by peer learning model on the University-Earth dataset. From left to right: ground-view query image and the Top 1-4 retrieved drone-view images. Green and red borders indicate correct and incorrect retrieved results, respectively. Note that peer learning model can find out the target sub-region.

B. Implementation Details

In PLCD, we adopt the ResNet-50 [43] pre-trained on ImageNet [44] as CNN$_1$, CNN$_2$, CNN$_3$ and CNN$_4$. We remove the last pooling layer and original classifier for ImageNet, then add one new classifier module. The new classifier module contains five layers: a max-pooling layer → a fully connected layer (FC$_1$) → a batch normalization layer (BN) → a dropout layer (Dropout) → a fully connected layer (FC$_2$). The max-pooling layer output 2048-dim feature and the FC$_1$ project it to 512-dim. All metric learning in our work are based on 512-dim representations. We resize each input image to a fixed size of 384 × 384 pixels during training and testing. We set $\tau = 0.1$, and $L = \{1, 2, 3, 4\}$ which means R-MAC layer can output the potential sub-regions with multi-scale 12×12, 9×9, 7×7 and 5×5. The model is trained by stochastic gradient descent with momentum 0.9. The learning rate is 0.01 for the classifier module and 0.001 for the rest layers. Dropout rate is 0.5. And λ_1 and λ_2 are all set to 1. We train our model for 120 epochs, and the learning rate is decayed by 0.1 after 40 epochs. During testing, we utilize the cosine similarity to measure the similarity between the query image and candidate images in the gallery.

C. Comparison with State-of-the-arts

Regarding to the ground → satellite retrieval, we compared our method, with several typical state-of-the-arts on the University-Earth and University-Google datasets respectively, including DEep Local Features (DELF) [36], R-MAC [41], Strong CNN features (Str-CNN) [10], CVM-Net [9], Siam-FCANet50 [12], LPN [26], Instance Loss [42], *etc.* Table. I shows the results. Note that the notation with a '*w/o* $\mathcal{D}$' suffix means that the indicated model has only been trained with the ground and satellite sets. If '*w/o* $\mathcal{D}$' is not included, it means that the model has been trained with ground, drone, and satellite sets. The results clearly demonstrate that our method PLCD has achieved 40.87%/21.77% CMC@1% accuracy and 14.16%/7.63% mAP on the University-Earth/University-Google dataset on ground-to-satellite (ground → satellite) image retrieval. The proposed method achieves the best performance and outperforms the state-of-the-arts with a very large margin, for example, our method is higher than the second best method (Str-CNN + Multi-loss) by 30.32%/14.64% and 11.04%/5.25% on CMC@1% and mAP respectively.

D. The effectiveness of Peer Learning

To evaluate the effectiveness of peer learning, we conduct some experiments on the University-Earth dataset. we first compared our method with state-of-the-arts on ground-to-drone image retrieval task (ground → drone), including DEep Local Features (DELF) [36], R-MAC [41], Strong CNN features (Str-CNN) [10], CVM-Net [9], and Siam-

TABLE II
THE COMPARISON WITH STATE-OF-THE-ARTS IN THE SETTING OF GROUND → DRONE RETRIEVAL.

Method	CMC@1	CMC@5	CMC@10	CMC@1%	mAP
DELF [36]	0.04	0.27	0.70	14.97	1.40
R-MAC [41]	2.09	3.99	5.12	21.79	2.34
Str-CNN [10]	1.51	3.26	4.96	18.65	1.65
Str-CNN+Semi-hard [38]	2.83	6.05	8.45	29.93	3.09
Str-CNN+Semi-hard+GeM [13]	2.44	5.51	7.95	30.21	3.04
CVM-Net [9]	2.79	5.12	7.02	27.18	2.59
Siam-FCANet50 [12]	3.14	4.89	6.48	24.89	3.07
Ours	**4.77**	**9.46**	**13.65**	**45.99**	**6.32**

TABLE III
ABLATION STUDY OF PEER LEARNING. THE TWO-BRANCH MODELS ARE THE BASE NETWORK. 'S' AND 'J' STAND FOR THE STEPS OF TRAINING THE SENIOR AND JUNIOR PEER, RESPECTIVELY. 'B' REPRESENTS SELECTING THE BEST SUB-REGION.

Method	CMC@1	CMC@5	CMC@10	CMC@1%	mAP
Two-branch	3.14	6.17	8.18	33.35	3.48
Two-branch+S	3.61	7.13	10.00	37.61	3.26
Two-branch+S+J	4.61	8.92	12.18	40.44	3.65
Two-branch+S+J+B	**4.77**	**9.46**	**13.65**	**45.99**	**6.32**

TABLE IV
COMPARISON OF DIFFERENT MAPPING METHODS BETWEEN GROUND-DRONE AND SATELLITE-DRONE DOMAINS.

Method	CMC@1	CMC@5	CMC@10	CMC@1%	mAP
Peer Learning	0.97	4.73	7.99	8.72	2.43
Projection	0.85	2,68	4.38	8.30	1.72
Supervision	2.36	7.08	11.69	12.34	3.82
Diffusion	**9.15**	**27.66**	**38.38**	**40.87**	**14.16**

TABLE V
COMPARISON OF ONE COMMON MODEL AND OUR TWO-BRANCH MODEL IN DIFFERENT SETTINGS.

Method	CMC@1	CMC@5	CMC@10	CMC@1%	mAP
Ground→Drone					
One Model	3.33	6.98	9.15	34.90	4.49
Two-branch	**4.77**	**9.46**	**13.65**	**45.99**	**6.32**
Drone→Satellite					
One Model	63.95	77.29	80.30	80.71	66.96
Two-branch	**67.06**	**82.84**	**87.23**	**87.69**	**70.60**
Ground→Satellite					
One Model	6.62	17.77	26.29	27.76	10.23
Two-branch	**9.15**	**27.66**	**38.83**	**40.87**	**14.16**

FCANet50 [12]. Table. II shows the results. The results demonstrate that our peer learning methods get the best performance compared with all other methods. It shows the effectiveness of the proposed peer learning network.

Second, we conduct an ablation study for the two steps of proposed peer learning. The base network for ground-drone representation exploits a structure of two-branch models. Senior and Junior are the two steps of peer learning strategy. B represents the feature of the best sub-region. It means that, for each gallery image, we use the feature of its best sub-region as its representation. We conduct experiments by adding them one by one. We show the ground-to-drone image retrieval performance in Table. III. The results show that each design for peer learning is very important and effective for peer learning.

Moreover, we show some retrieval results on Figure 5. We observe that our model can retrieve the reasonable image and its sub-region based on the content. Some failure cases are also shown in the fourth column of drone-view image list. We notice that it is challenging in that the irrelevant drone image has a similar pattern with the ground image.

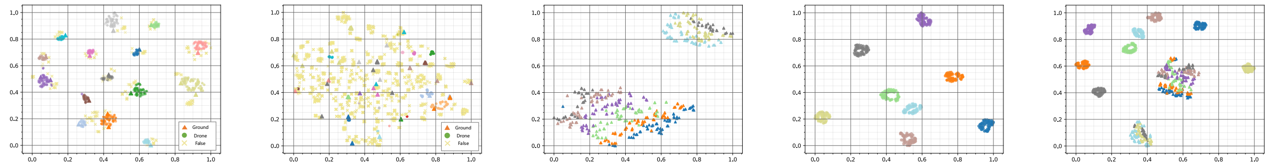
To provide more evidence of the effectiveness of our method, we also visualize the feature distributions. We sample some ground-view images in the testing set of the University-Earth dataset and select their top-50 retrieval results. Figure 6 (a)-(b) show the visualization distribution. Compared with Str-CNN [10]+Multi-loss, the distributions show our peer learning can achieve a better distribution on ground-drone cross-view space and find more positives.

E. The effectiveness of Cross Diffusion

To show the effectiveness of the proposed diffusion methods. We compared with different mapping methods to connect the ground-drone and satellite-drone domains. The mapping methods include: 1) Peer Learning: applying peer learning strategy to train a common space for ground-drone-satellite representation; 2) Projection: training a new fully connective (FC) layer to project drone-view features of ground-drone and satellite-drone models to a same space; 3) Supervision: when training the drone-satellite model, for each drone-view image, we close the distance between its features extracted by the drone-satellite model and the pretrained ground-drone model; 4) Diffusion: the proposed cross diffusion method. The the ground-satellite image retrieval performances by these four methods are shown in Table. IV. The results clearly indicate that our diffusion method is the most effective method to connect the ground-drone and satellite-drone domains.

We sample some drone-view images from University-Earth's testing set and visualize their feature distribution in ground-drone (peer learning model) and satellite-drone (patch-based model) spaces. Then we project all features into the same space called ground-drone-satellite space. Figure 6 (c)-(e) show the results. We can find that, in the ground-drone and satellite-drone spaces, features of each identity are relatively grouped in clusters. But even we try to project drone-view features from these two spaces to the same space, most of the samples are off the manifolds of their corresponding classes. This supports our assumption that mapping images from different space into a uniform space distorts the manifolds.

Note that cross diffusion requires the creation of graphs in two independent spaces. The quality of the graphs affects the final diffusion result, and peer learning can help us build better mapping spaces and thus improve the quality of the graphs, so peer learning is not a redundant part. We also apply diffusion on Str-CNN + Multi-loss and CVM-Net, achieved 28.29% CMC@1% accuracy, 10.03% mAP and 26.18% CMC@1% accuracy, 9.21% mAP respectively on ground-drone-satellite image retrieval (on the University-Earth dataset), all less accurate than PLCD. Hence, Table. IV also shows that simply building a connection between two independent spaces does



(a) (b) (c) (d) (e)

Fig. 6. Visualization of feature distributions. (a) and (b) show the feature distributions of peer learning and Str-CNN [10], respectively. For each ground-view image (donated by the triangle), we select its top-50 retrieval results from drone-view gallery list. Positive and negative results are donated as circles and × marks, respectively. (c), (d), (e) show the feature distributions of ground-drone, satellite-drone, and ground-drone-satellite spaces, respectively. Features from the ground-drone and satellite-drone are denoted as triangles and circles, respectively.

not achieve good results and also illustrates the effectiveness of diffusion.

F. How important are the Drone-view Images?

To demonstrate that drone-view images can facilitate ground-to-satellite image retrieval task, we select state-of-the-art methods and train them on the ground and satellite sets, or on the ground, drone, and satellite sets. The results are shown in Table. I. In the table, notation with a '*w/o*' suffix means that the indicated model is only trained on the ground and satellite sets without drone-view images. The results show that every method can improve its performance by adding drone-view images for training. The performance demonstrates a fact that drone-view information benefits the ground-to-satellite image retrieval task, and plays an important role to make a bridge between ground-view and satellite-view images.

In our method, drone images (**no label**) are set as the reference set at query time. The content overlap between ground and satellite images of the same landmark is very small, it is pretty difficult to extract relevant information between them without the drone images. And due to the diversity of landmarks, even if drone images are used in training, it is difficult for the model to make correct inferences at query time if drone images are not used just as a reference.

Here we would like to discuss the practicality of using drone data as a reference. In the real world, it would be really difficult to collect drone images in some cities, but we can use synthetic data to solve this problem. In the dataset we used, all the ground images and satellite images are collected in the real world, while the drone images are **synthetic data** collected from google-earth. From the experiments, we can see that the synthetic drone images help to improve the ground-stallite retrieval task. In addition, in the test, the drone data as reference does not need any annotation, so in the real application, we only need to provide the model with the synthetic data of the corresponding area. We believe our task fits the realistic application well.

G. Why do not use One Model for All View?

The viewpoint varies greatly between ground-view and satellite-view/drone-view images. We consider it's hard to design a common representation model for images of all views. To prove this, we make a comparison of one common model and the two-branch model in the setting of Ground→Drone retrieval, Drone→Satellite retrieval and Ground→Satellite retrieval. Table. V shows the results of all settings trained on the University-Earth dataset. The results clearly demonstrate that the two-branch structure outperforms the one common model in all settings. Thus, a two-branch structure is more suitable for this kind of ground-drone-satellite retrieval tasks that have large viewpoint changes.

H. Experiments on the parameters

The τ is the temperature hyper-parameter. When the value of τ less than 1, it makes the similarity vector sharper and vice versa. We do not expect that the distribution space of the junior model is too similar to the senior model. We consider that making the similarity vector sharper can improve the power of supervision from the senior model and achieves better results.

To support our assumption, we conducted experiments with different values of τ ($\tau = 2, 0.5, 0.1, 0.05$ and 0.01) on the University-Earth dataset. Figure 7 reports the results. The results show that our model achieves the best result when $\tau = 0.1$.

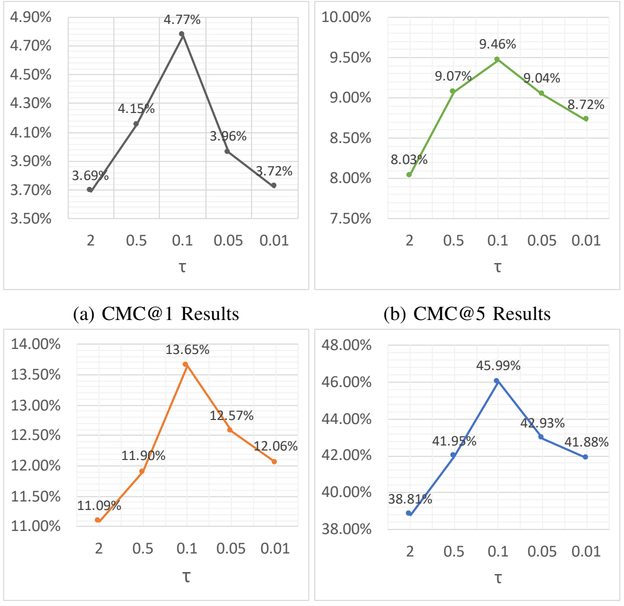
We also have experiments on the parameters λ_1 and λ_2. The results are shown in Table VI, which indicate that our model achieves the best result when $\lambda_1 = 1$ and $\lambda_2 = 1$.

λ_1	CMC@1%	mAP (%)
0.1	42.65	5.36
0.5	43.62	6.05
1	**45.99**	**6.32**
2	41.99	5.86
5	33.89	4.12
λ_2	CMC@1%	mAP (%)
0.1	80.18	64.64
0.5	86.10	62.25
1	**87.69**	**70.60**
2	83.72	63.37
5	79.89	54.12

TABLE VI
RESULTS ON DIFFERENT PARAMETERS λ_1 AND λ_2.

I. The number of training samples

We conducted experiments on the University-Earth dataset by using different batchsize. The batchsize N means that we randomly select N street/satellite images and $6N$ drone images in each mini-batch. The results of different batchsize are shown in Table VII. Hence, considering the batchsize for the best performance, we set the batchsize to 8/4 in our work.



(a) CMC@1 Results (b) CMC@5 Results

(a) CMC@10 Results (b) CMC@1% Results

Fig. 7. Results on different parameter τ.

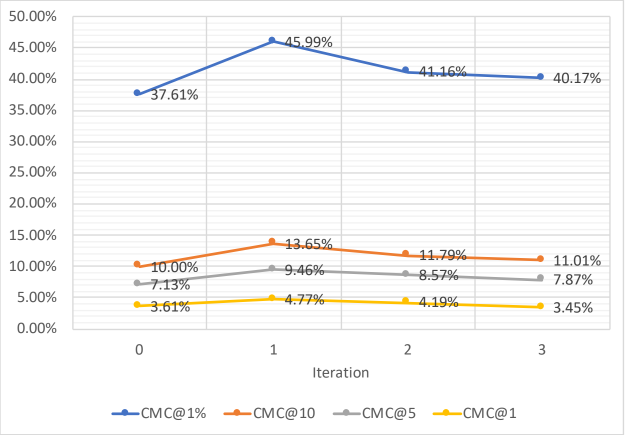


Fig. 8. Results on different iterations ($\tau = 0.1$).

N	CMC@1%	mAP (%)
4	30.24	2.77
6	39.32	4.78
8	**45.99**	**6.32**
12 (two GPUs)	41.84	4.76
16 (two GPUs)	41.57	4.49

N	CMC@1%	mAP (%)
2	71.71	52.54
4	**87.69**	**70.60**
6	86.51	68.47
8	84.17	63.91
10	80.73	59.83

TABLE VII
RESULTS ON DIFFERENT BATCHSIZE N FOR GROUND-DRONE REPRESENTATION AND DRONE-SATELLITE REPRESENTATION, RESPECTIVELY.

J. Experiments on the iteration of peer learning

Soft label estimation has been widely studied in self-supervised learning methods, where the network predicts soft labels by properly utilizing the capability of the network itself. Generally speaking, the model is trained in several runs via self-predicted soft labels. Some previous works [45], [46] demonstrate the effectiveness of progressive learning in iteration.

To investigate the effectiveness of peer learning in iteration via soft labels, we conducted experiments on the University-Earth dataset with different iterations. In each iteration, we use the junior model from the last iteration as the senior model to finish the peer learning process. We set $\tau = 0.1$. Figure 8 shows the results. The results show that our model achieves the best result in the first iteration.

K. Visualization Results and Failure Case

In this subsection, we visualize more retrieval results. Given a ground-view image, we show its top five drone-view images by Ground→Drone retrieval, as well as top five satellite-view images by Ground→Satellite retrieval. The visualization results are shown in Figure 9. The top three rows show some good results, we can see that although the viewpoint changes a lot and the landmark is not so obvious in the images, our method can find the satellite images in the top results.

The bottom two rows also show two failure cases. For the fourth query, we can find some drone-view images, but we fail in finding its satellite-view images. We consider it is because the landmark is a little small and common and does not contain too many special details in the satellite-view images. For the fifth query, the landmark is relatively not so obvious in the ground-view image, making it hard to be found in the drone-view images.

Furthermore, we visualize more retrieval results of the hard samples. Given a hard sample (ground-view image), we show its top five drone-view images by Ground→Drone retrieval, as well as top-five satellite-view images by Ground→Satellite retrieval. The visualization results are shown in Figure 10. The top three rows show some results of occlusion, and the bottom three rows show some results of tiny parts. The results demonstrate that our method can be effective in hard samples.

L. Computational Efficiency

For the efficiency, when we train on the University-Earth dataset (2,659 street images, 37,854 drone images and 701 satellite images, single Tesla V100S GPU), the training time of the first step Ground-Drone Cross-view Representation and the second step Satellite-Drone Cross-view Representation are $731m\pm3m38s$ (Senior Peer: $323m18s\pm1m38s$; Junior Peer: $403m\pm2m$) and $87m11s\pm11s$, respectively.

When we test on the University-Earth dataset (2,579 query-street images, 17,119 gallery-drone images and 951 satellite images), the running time of extracting feature and cross-diffusion ranking are $5m26s\pm5s$ ($37s$ for query images and $4m49s\pm5s$ for gallery images, single Tesla V100S GPU) and $146s\pm2s$ (on CPU), respectively. For each query, the running

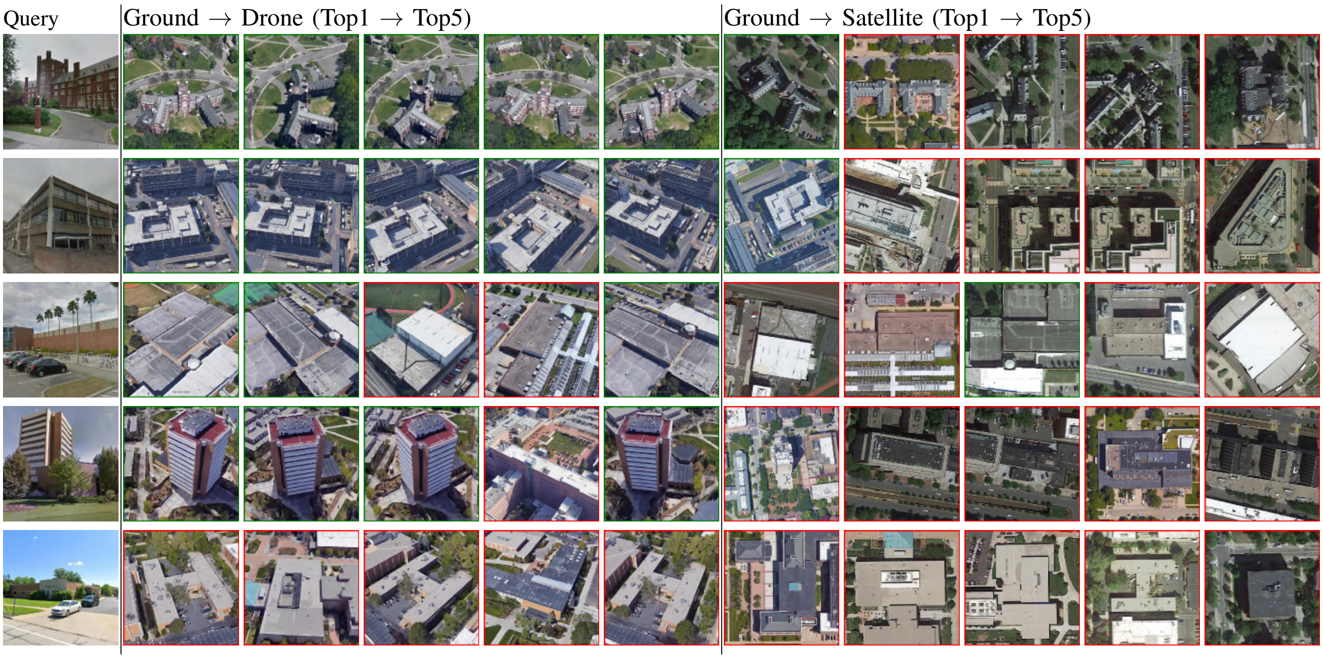

Query
Ground → Drone (Top1 → Top5)
Ground → Satellite (Top1 → Top5)

Fig. 9. The visualization of ground-to-satellite image retrieval by PLCD framework on University-Earth dataset. There are two results, from left to right: ground-view query image, the Top 1-5 retrieved drone-view images and the Top 1-5 retrieved satellite-view images. Green and red borders indicate correct and incorrect retrieved results, respectively.

time of our method is $0.18s$ in average. All information are shown in Table VIII. Hence, our method works in real-time application.

V. Conclusion

In this paper, we argue the defects/challenges of existing approaches in ground-to-satellite geo-localization task, and raise a new cross-view representation and diffusion strategy. The key novelty of the paper is the idea of exploiting drone-view images in training to improve the performance of ground-satellite image retrieval. As minor contributions, our techniques are also of sufficient novelty. 1) For the framework, we have two independent cross-view networks; 2) For the network architecture, to address the specific mismatch challenge in cross-view image retrieval, by adopting R-MAC and detection layer; 3) For optimization, we adopt peer learning strategy, which is firstly applied to the cross-view image retrieval task; 4) Our diffusion works as a connection between two independent feature spaces, while existing diffusion just runs in single feature space.

Perspective Drone images are not as easy to locate less accurately, compared to satellite images, hence we use satellite images to achieve geo-localization. Meanwhile, the drone-view images for our PLCD do not need annotations or any preparation according to the query. We believe our task fits the realistic application well. We find that even if we apply a network that achieves a better performance on drone-to-satellite retrieval, it may not necessarily obtain better results on diffusion. What is the relevance between cross-diffusion and cross-view space? We will continuously investigate this point worth nothing. In addition, the external environment also has a significant impact on the ground-to-satellite geo-localization. Small targets [47] and retrieval in low light [48], [49], [50], [51] are also worth investigating.

References

[1] Y. Zhang, X. Qian, X. Tan, J. Han, and Y. Tang, "Sketch-based image retrieval by salient contour reinforcement," *IEEE Trans. Multimedia*, vol. 18, no. 8, pp. 1604–1615, 2016.

[2] Y. He, S. Xiang, C. Kang, J. Wang, and C. Pan, "Cross-modal retrieval via deep and bidirectional representation learning," *IEEE Trans. Multimedia*, vol. 18, no. 7, pp. 1363–1377, 2016.

[3] X. Gu, Y. Wong, L. Shou, P. Peng, G. Chen, and M. S. Kankanhalli, "Multi-modal and multi-domain embedding learning for fashion retrieval and analysis," *IEEE Trans. Multimedia*, vol. 21, no. 6, pp. 1524–1537, 2018.

[4] X. Lu, L. Zhu, J. Li, H. Zhang, and H. T. Shen, "Efficient supervised discrete multi-view hashing for large-scale multimedia search," *IEEE Trans. Multimedia*, vol. 22, no. 8, pp. 2048–2060, 2019.

[5] C. Zheng, L. Zhu, Z. Cheng, J. Li, and A. Liu, "Adaptive partial multi-view hashing for efficient social image retrieval," *IEEE Trans. Multimedia*, 2020.

[6] A. Verma, A. Subramanyam, Z. Wang, S. Satoh, and R. R. Shah, "Unsupervised domain adaptation for person re-identification via individual-preserving and environmental-switching cyclic generation," *IEEE Transactions on Multimedia*, 2021.

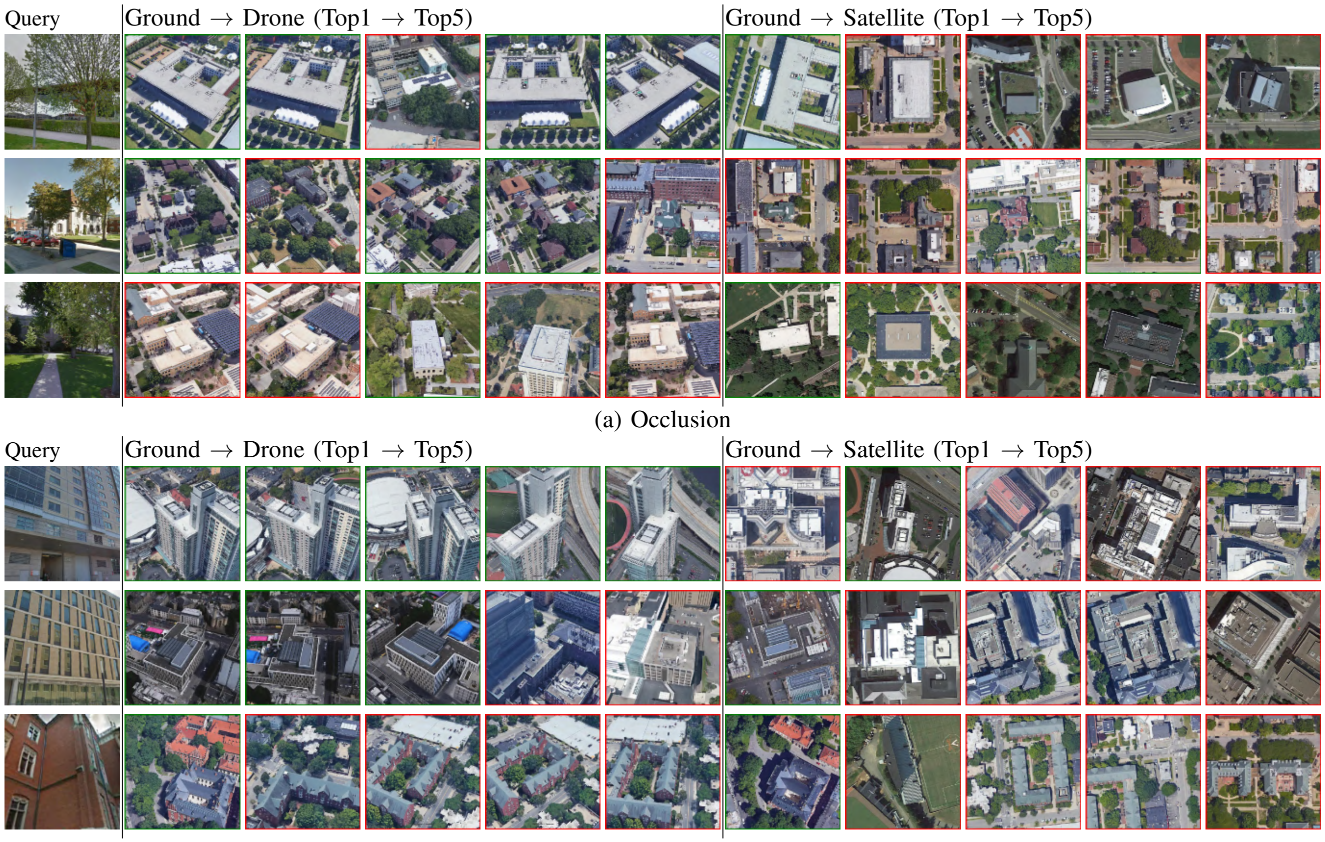
[7] S. Workman, R. Souvenir, and N. Jacobs, "Wide-area image geolocalization with aerial reference imagery," in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 3961–3969.

[8] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5297–5307.

[9] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7258–7267.

[10] Z. Zheng, Y. Wei, and Y. Yang, "University-1652: A multi-view multi-source benchmark for drone-based geo-localization," in *Proc. ACM Int. Conf. Multimedia*, 2020.

[11] N. N. Vo and J. Hays, "Localizing and orienting street views using overhead imagery," in *Proc. European Conf. Comput. Vis.*, 2016, pp. 494–509.


Query
Ground → Drone (Top1 → Top5)
Ground → Satellite (Top1 → Top5)
Query
Ground → Drone (Top1 → Top5)
Ground → Satellite (Top1 → Top5)

(a) Occlusion

(b) Tiny Part

Fig. 10. The visualization of the ground-to-satellite retrieval results of hard samples by our PLCD framework, including (a) occlusion samples and (b) tiny part samples. From left to right: ground-view query image, the Top 1-5 retrieved drone-view images, and the Top 1-5 retrieved satellite-view images. Green and red borders indicate correct and incorrect retrieved results, respectively.

Part		Time (single Tesla V100S GPU)
Ground-Drone Cross-view Representation	Senior Peer	Training time: $323m18s\pm1$m38s
	Junior Peer	Training time: $323m18s\pm1m38s$
	Total	Training time: $403m\pm2m$
Satellite-Drone Cross-view Representation		Training time: $87m11s\pm11s$
Cross-View Image Retrieval	Feature Extraction	Test time: $5m26s\pm5s$
	Cross-diffusion Ranking	Testing time: $146s\pm2s$ (CPU)
	Average	Testing time: 0.18s per query

TABLE VIII
RESULTS OF THE TRAINING AND TESTING TIME OF PLCD.

[12] 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 *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 8391–8400.

[13] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning cnn image retrieval with no human annotation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1655–1668, 2018.

[14] Y. Shi, X. Yu, L. Liu, T. Zhang, and H. Li, "Optimal feature transport for cross-view image geo-localization." in *Proc. AAAI Conf. Artificial Intell.*, 2020, pp. 11 990–11 997.

[15] S. Bai, Z. Zhou, J. Wang, X. Bai, L. Jan Latecki, and Q. Tian, "Ensemble diffusion for retrieval," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 774–783.

[16] S. Bai, X. Bai, Q. Tian, and L. J. Latecki, "Regularized diffusion process on bidirectional context for object retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 5, pp. 1213–1226, 2018.

[17] S. Hu and X. Chang, "Multi-view drone-based geo-localization via style and spatial alignment," in *Proc. ACM Int. Conf. Multimedia*, 2020.

[18] 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.*, 2019, pp. 10 090–10 100.

[19] L. Liu and H. Li, "Lending orientation to neural networks for cross-view geo-localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5624–5633.

[20] S. Workman, R. Souvenir, and N. Jacobs, "Wide-area image geolocalization with aerial reference imagery," in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 1–9.

[21] M. Zhai, Z. Bessinger, S. Workman, and N. Jacobs, "Predicting ground-level scene layout from aerial imagery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 867–875.

[22] W. Huang, R. Hu, X. Wang, C. Liang, and J. Chen, "Occluded suspect search via channel-guided mechanism," *Neural Computing and Applications*, vol. 33, no. 3, pp. 961–971, 2021.

[23] J. I. Olszewska, "Interest-point-based landmark computation for agents' spatial description coordination." in *ICAART (2)*, 2016, pp. 566–569.

[24] ——, "Clock-model-assisted agent's spatial navigation." in *ICAART (2)*, 2017, pp. 687–692.

[25] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. European Conf. Comput. Vis.*, 2018, pp. 480–496.

[26] T. Wang, Z. Zheng, C. Yan, J. Zhang, Y. Sun, B. Zhenga, and Y. Yang,

"Each part matters: Local patterns facilitate cross-view geo-localization," *IEEE Trans. Circuits Syst. Video Technol.*, 2021.

[27] X. Zhong, T. Lu, W. Huang, M. Ye, X. Jia, and C.-W. Lin, "Grayscale enhancement colorization network for visible-infrared person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[28] F. Yang, Y. Wu, Z. Wang, X. Li, S. Sakti, and S. Nakamura, "Instance-level heterogeneous domain adaptation for limited-labeled sketch-to-photo retrieval," *IEEE Transactions on Multimedia*, 2020.

[29] K. Kansal, A. V. Subramanyam, Z. Wang, and S. Satoh, "Sdl: Spectrum-disentangled representation learning for visible-infrared person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 10, pp. 3422–3432, 2020.

[30] N. Vo, L. Jiang, C. Sun, K. Murphy, L.-J. Li, L. Fei-Fei, and J. Hays, "Composing text and image for image retrieval-an empirical odyssey," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6439–6448.

[31] X. Guo, H. Wu, Y. Cheng, S. Rennie, G. Tesauro, and R. S. Feris, "Dialog-based interactive image retrieval," *arXiv preprint arXiv:1805.00145*, 2018.

[32] K. Pang, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Cross-domain generative learning for fine-grained sketch-based image retrieval." in *BMVC*, 2017, pp. 1–12.

[33] J. Song, Q. Yu, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Deep spatial-semantic attention for fine-grained sketch-based image retrieval," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5551–5560.

[34] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[35] A. Gordo, J. Almazán, J. Revaud, and D. Larlus, "Deep image retrieval: Learning global representations for image search," in *Proc. European Conf. Comput. Vis.* Springer, 2016, pp. 241–257.

[36] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han, "Large-scale image retrieval with attentive deep local features," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 3456–3465.

[37] L. Liu, H. Li, and Y. Dai, "Stochastic attraction-repulsion embedding for large scale image localization," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 2570–2579.

[38] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 815–823.

[39] F. Yang, Z. Wang, and J. Xiao, "Mining on heterogeneous manifolds for zero-shot cross-modal image retrieval," in *Proc. AAAI Conf. Artificial Intell.*, 2020.

[40] J. Toman and J. Olszewska, "Algorithm for graph building based on google maps and google earth," in *2014 IEEE 15th International Symposium on Computational Intelligence and Informatics (CINTI)*. IEEE, 2014, pp. 55–60.

[41] G. Tolias, R. Sicre, and H. Jégou, "Particular object retrieval with integral max-pooling of cnn activations," in *ICL 2016-RInternational Conference on Learning Representations*, 2016, pp. 1–12.

[42] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, and Y.-D. Shen, "Dual-path convolutional image-text embeddings with instance loss," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 16, no. 2, pp. 1–23, 2020.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[44] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[45] T. Furlanello, Z. C. Lipton, M. Tschannen, L. Itti, and A. Anandkumar, "Born-again neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2018.

[46] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10 687–10 698.

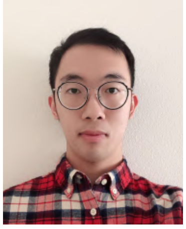
[47] M. Hu, J. Xiao, L. Liao, Z. Wang, C.-W. Lin, M. Wang, and S. Satoh, "Capturing small, fast-moving objects: Frame interpolation via recurrent motion enhancement," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[48] Z. Zeng, Z. Wang, Y. Zheng, Y.-Y. Chuang, and S. Satoh, "Illumination-adaptive person re-identification," *IEEE Transactions on Multimedia*, vol. 22, no. 12, pp. 3064–3074, 2020.

[49] K. Jiang, Z. Wang, P. Yi, C. Chen, Z. Han, T. Lu, B. Huang, and J. Jiang, "Decomposition makes better rain removal: An improved attention-guided deraining network," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

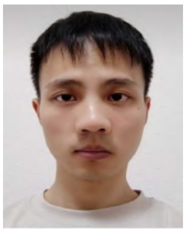
[50] K. Jiang, Z. Wang, P. Yi, C. Chen, Z. Wang, X. Wang, J. Jiang, and C.-W. Lin, "Rain-free and residue hand-in-hand: A progressive coupled network for real-time image deraining," *IEEE Transactions on Image Processing*, vol. 30, pp. 7404–7418, 2021.

[51] X. Xu, S. Wang, Z. Wang, X. Zhang, and R. Hu, "Exploring image enhancement for salient object detection in low light images," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 17, no. 1s, pp. 1–19, 2021.



Zelong Zeng is currently a PhD student at the Department of Information and Communication Engineering, Graduate School of Information Science and Technology, the University of Tokyo, Tokyo, Japan. He receive the M.S degrees from the University of Tokyo in 2020. His research interests include person re-identification and image retrieval.

Zheng Wang (M'19) received the B.S., M.S., and Ph.D. degrees from Wuhan University in 2006, 2008, 2017, respectively. He was a JSPS Fellowship Researcher at the National Institute of Informatics, Japan, and a Project Assistant Professor at The University of Tokyo, Japan. He is currently a Professor at Wuhan University, China. His research interests include person re-identification and instance search.



Fan Yang received the B.E. degree from Zhejiang University in 2015, and the M.E. and Ph.D. degrees from the University of Tokyo in 2018 and 2021, respectively. He is currently a researcher at the National Institute of Informatics, Japan. His research interests include image/video retrieval and person re-identification.

Shin'ichi Satoh (M'04) received the B.E. degree in electronics engineering and the M.E. and Ph.D. degrees in information engineering from the University of Tokyo, Tokyo, Japan, in 1987, 1989, and 1992, respectively. He has been a Full Professor with the National Institute of Informatics, Tokyo, Japan, since 2004. He was a Visiting Scientist with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA, from 1995 to 1997. His current research interests include image processing, video content analysis, and multimedia databases.