It's Okay to Be Wrong: Cross-View Geo-Localization With Step-Adaptive Iterative Refinement

Xiufan Lu, Siqi Luo, and Yingying Zhu, *Member, IEEE*

***Abstract*— Cross-view image geo-localization is a challenging task of estimating the geospatial location of a street-view image by matching it with a database of geotagged aerial/satellite images, and vice versa. Compared to existing CNN-based approaches that attempt to generate discriminative representations in a single step for this task, in this article, we instead advocate endowing the network with the capability of progressive self-correcting. Toward this target, we propose a novel step-adaptive iterative refinement network (SIRNet), which decomposes the complex learning process into several refinement steps while adapting the refinement steps specifically for each input. Specifically, the SIRNet takes the output of the backbone as a rough network prediction and iteratively refines it via an iterative refinement module (IRM). The IRM cascades several refinement blocks sharing the same structure for progressive self-correcting. For each refinement block, the goal is to improve the output of the previous refinement block under the guidance of height-wise context. In this way, the IRM is capable of improving the rough network prediction step by step, and the refined features are increasingly focused on more discriminative scene regions as they are iteratively refined. In addition, considering different characteristics of input images, we devise an adaptive step estimation (ASE) mechanism, which enables our SIRNet to adapt the number of refinement steps to each input automatically. Concretely, the ASE is performed by comparing features at adjacent refinement steps, estimating whether the next step brings improvements, and finally making a halting decision at each refinement step. With the ASE, our SIRNet becomes a dynamic architecture that considers different characteristics of the inputs when performing the iterative refinement. Extensive experiments demonstrate that our SIRNet performs favorably against the state-of-the-art methods on the CVUSA and the CVACT datasets. Furthermore, quantitative and qualitative experimental results demonstrate our approach's wide applicability, impressive generalization ability, and robustness.**

***Index Terms*— Adaptive estimation, convolutional neural network, cross-view geo-localization, image retrieval, iterative refinement.**

I. Introduction

With the development of aerospace technology and sensor technology, researchers can easily acquire large amounts of high-quality remote sensing images, which reflect the state of the ecological environment and traces of human activities [1]. This provides a new clue to solving the image-based geo-localization, which aims to estimate the geo-location of an image by comparing it against geo-tagged database images [2], [3]. Recently, cross-view geo-localization (also known as ground-to-aerial geo-localization) has become an attractive proposition for addressing the image-based geo-localization. Specifically, as depicted in Fig. 1, cross-view geo-localization aims to determine the location of a street-view image by matching it with a series of geo-tagged aerial images covering the same or wider region, and vice versa. Cross-view geo-localization is significantly challenging for two main reasons. On the one hand, viewpoints between ground and aerial images change drastically, which results in significant appearance and geometric differences between ground and aerial images. On the other hand, there are visual interferences in ground images or aerial images, such as variable illumination and transient occlusions (e.g., pedestrian and cars). To overcome these difficulties, existing works typically treat this task as an image-retrieval task, whose key is to generate discriminative representations to distinguish between similar-looking locations.

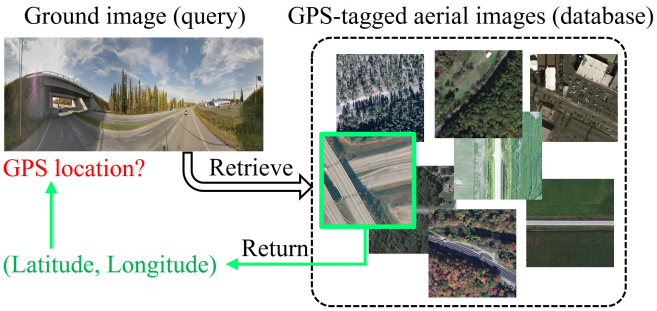

Fig. 1. Demonstration of cross-view geo-localization.

Significant efforts have been made by incorporating attention mechanisms [5], [7], specialized loss functions [7], [8], [9], [10], [11], contextual information [12], or orientation knowledge [6], [13]. However, as shown in Fig. 2, the state-of-the-art methods still struggle to recognize discriminative regions accurately. This makes us recognize that it remains difficult to infer discriminative regions in a single forward pass due to large viewpoint variance and visual interferences. By contrast, humans are better at handling complex tasks in steps and good at self-reflection and self-correcting. Inspired by this observation, we propose a novel step-adaptive iterative

Manuscript received 21 April 2022; revised 23 June 2022 and 13 August 2022; accepted 20 August 2022. Date of publication 10 October 2022; date of current version 9 November 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62072318, in part by the Natural Science Foundation of Guangdong Province of China under Grant 2021A1515012014, in part by the Science and Technology R&D Funds of Shenzhen under Grant JCYJ20190808172007500 and Grant 20220810142553001, and in part by the China University Industry-Academia-Research Innovation Funds under Grant 2021LDA12014. *(Corresponding author: Yingying Zhu.)*

The authors are with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, Guangdong 518052, China (e-mail: zhuyy@szu.edu.cn).

Digital Object Identifier 10.1109/TGRS.2022.3210195

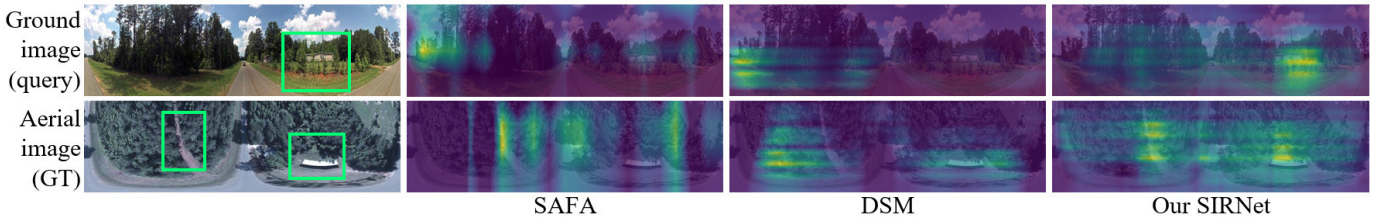

Fig. 2. Visualization of generated feature maps on the CVUSA dataset [4] comparing our proposed SIRNet with two state-of-the-art models, i.e., SAFA [5] and DSM [6]. For a fair comparison, we warp aerial images using polar transform as in [5] and [6], a simple trick that geometrically aligns ground and aerial images. "GT" is short for ground truth. We indicate regions with higher activation values in yellow, and for ease of reference, we box the discriminative scene regions in green.

refinement network (SIRNet) capable of progressive self-correcting. Our method iteratively recognizes discriminative regions, given the rough prediction output from the backbone network as priors. Since the rough prediction is an important clue about the rough position of discriminative regions, learning to iteratively refine the predicted results enables the network to leverage knowledge in previous iterations. As a result, the network is allowed to "make mistakes" (i.e., focusing on undiscriminating regions) at primary stages, without having to make accurate predictions in a single shot, as long as it can gradually adjust and correct these mistakes. As shown in Fig. 2, our network incorporated with the iterative refinement mechanism exceeds in recognizing discriminative regions.

To support the iterative self-correction, we need to address two key issues. First, which cues should be organized to assist in self-correction, and how do they improve the rough prediction? Second, as we mentioned above, the self-correction is performed step by step, so how to determine the number of self-correction steps? In response to the first problem, we devise an iterative refinement module (IRM) with a cascade of refinement blocks (each for a single refinement step) sharing the same structure without sharing their weights. At each refinement step, the refinement block aggregates height-wise contexts that indicate the context of horizontally divided feature regions to assist in self-correction. Afterward, the context-guided self-correction is performed by reconsidering which features are more discriminative than others and refining the input features accordingly. With the IRM, the SIRNet is capable of improving predicted results at the region level, given only image-level supervision. For the second problem, we further propose to adapt the iterative refinement steps to each input sample. We empirically find that using static iterative refinement steps, i.e., applying the fixed number of refinement steps regardless of inputs will limit the model's performance because each image has its characteristics. Therefore, we propose an adaptive step estimation (ASE) mechanism to configure the refinement steps conditioned on each input. Specifically, we make a halting decision after each refinement step at test time. The SIRNet terminates the refinement process once a refined feature map is discriminative enough for cross-view geo-localization task (judged by a softmax confidence). Otherwise, the network continues the refinement until the maximum number of refinement steps is reached. Two examples depict this procedure in Fig. 3. Our SIRNet incorporated with the IRM and the ASE is capable of recognizing discriminative regions while suppressing visual interferences step by step.

To summarize, our contributions are threefold.

1) Instead of directly learning representations in a single step, we devise an iterative refinement approach, i.e., the IRM, which endows the network with the capability of progressive self-correcting. The IRM takes the output of the backbone as a rough prediction and refines it iteratively under the guidance of scene context. Due to its iterative and self-correcting nature, our proposed SIRNet can progressively highlight more discriminative scene regions with only image-level supervision.
2) We devise an ASE mechanism that adaptively estimates the number of iterative refinement steps specifically for each input sample. With the ASE, our SIRNet becomes a dynamic architecture that considers different characteristics of the input when performing iterative self-correcting, thus gaining remarkable representation capability.
3) We compare the SIRNet with the state-of-the-art methods on various cross-view geo-localization tasks, including standard, fine-grained and few-shot cross-view geo-localization. Quantitative and qualitative ablation studies demonstrate the advantages of our proposed method in terms of effectiveness and generalization ability.

II. Related Works

A. Cross-View Geo-Localization

Finding feature correspondence between ground and aerial images is extremely difficult due to large domain gap across views and visual interferences. Existing methods developed for this task can be divided into two categories.

The main idea of the first category is to bridge the cross-view domain gap. For example, CVM-Net [8], [14] employs a NetVLAD technique [3] to project cross-view features into a shared space where they are comparable. CVFT [15] achieves this goal by applying an optimal transport algorithm to transport ground features to aerial feature space. Although promising, the CVM-Net and the CVFT bridge the domain gap purely based on image content without considering geometric priors between ground and aerial views. To address this problem, Polar-SAFA [5] and DSM [6] use a

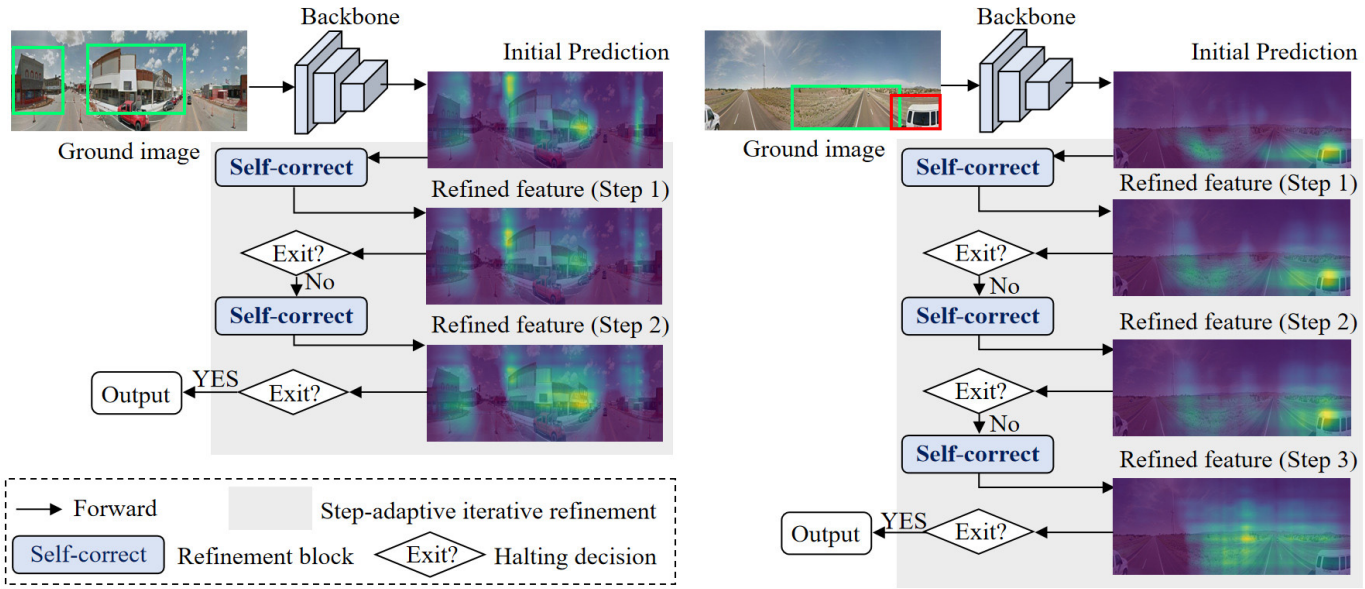

Fig. 3. Our proposed SIRNet iteratively refines the rough predictions output from the backbone with refinement steps conditioned on each input. (Left) Two-step iterative refinement process. (Right) Three-step iterative refinement process. The visualization examples demonstrate that the SIRNet is capable of highlighting discriminative regions (left and right) while suppressing transient occlusions (right) step by step. For ease of reference, we box discriminative regions and transient occlusions in green and red, respectively.

polar transform algorithm to warp aerial images so that they are geometrically aligned with ground images. Nevertheless, the polar transform does not consider the scene content and therefore cannot fully align the two domains. As a result, existing methods [5], [6] still have difficulty in recognizing discriminative scene objects, even if they apply the polar transform, as shown in Fig. 2. To achieve both semantic and geometric alignment, [16] and [17] introduce conditional GANs [18] to synthesize a corresponding aerial image from a ground query (or vice versa). However, since the cross-view synthesis [19], [20] is an ill-posed problem and is highly challenging, the synthesized images are always granulated and lack details, which hinders the domain alignment. This article does not intend to push complete alignment of the cross-view domains but to find feature correspondence by recognizing discriminative scene objects. Our method, therefore, can be well combined with the first category of works. As the experiments show, incorporating our proposed module with the first category of method consistently improves their performance.

The goal of the second category is to generate representations that are discriminative enough to distinguish between similar-looking images. Some efforts have been made toward this goal by designing specialized metric learning techniques to train models developed for cross-view geo-localization. Hu et al. [8] develop a weighted soft-margin loss based on the triplet loss [21], [22] to speed up training convergence. Then a hard sample mining mechanism [9], [10], [11] and a binomial loss [11] are further introduced to improve the performance of the weighted soft-margin loss. To effectively locate hard examples, Cai et al. [7] propose a hard exemplar reweighting loss that adaptively allocates different weights to triplets based on their difficulty. Recently, Rodrigues and Tani [23] design a data augmentation technique to produce more training samples for this task by keeping or removing scene objects based on their segmentation masks.

In addition to specialized metric learning techniques, several powerful CNN-based models are developed. Sun et al. [9] and Zhu et al. [10] cascade a ResNetX backbone with a capsule network [24] to model spatial feature hierarchies and enhance representation capability. Inspired by geometric cues, Liu and Li [13] incorporate orientation embeddings to CNN, endowing the network with the concept of orientation. However, this method requires the orientation knowledge to be provided, which may not be available in practice. To jointly obtain orientation and localization information, Shi et al. [6] propose a dynamic similarity matching method by sliding ground features along aerial features. Despite the effectiveness of these orientation-based methods, their performances are still limited when the network fails to focus on discriminative scene objects. To overcome this problem, SAFA [5] introduces a multihead spatial attention method to highlight salient scene regions. Cai et al. [7] further explore both spatial and channel attention for better performance. Nevertheless, we empirically observe that these attention-based methods struggle to grasp discriminative features using a single forward pass due to the large viewpoint variance. In contrast, our SIRNet adopts an iterative self-correcting scheme, which decomposes the learning process into multiple refinements and enables the network to make more accurate predictions step by step. Recently, Wang et al. [12] adopts a square-ring feature partition strategy to take advantage of contextual information for geo-localization. However, this method simply aggregates all information of neighbor areas as an image representation, which may introduce noises. Contrastively, our SIRNet aggregates height-wise context for progressive self-correcting, which exploits contextual information while highlighting discriminative features.

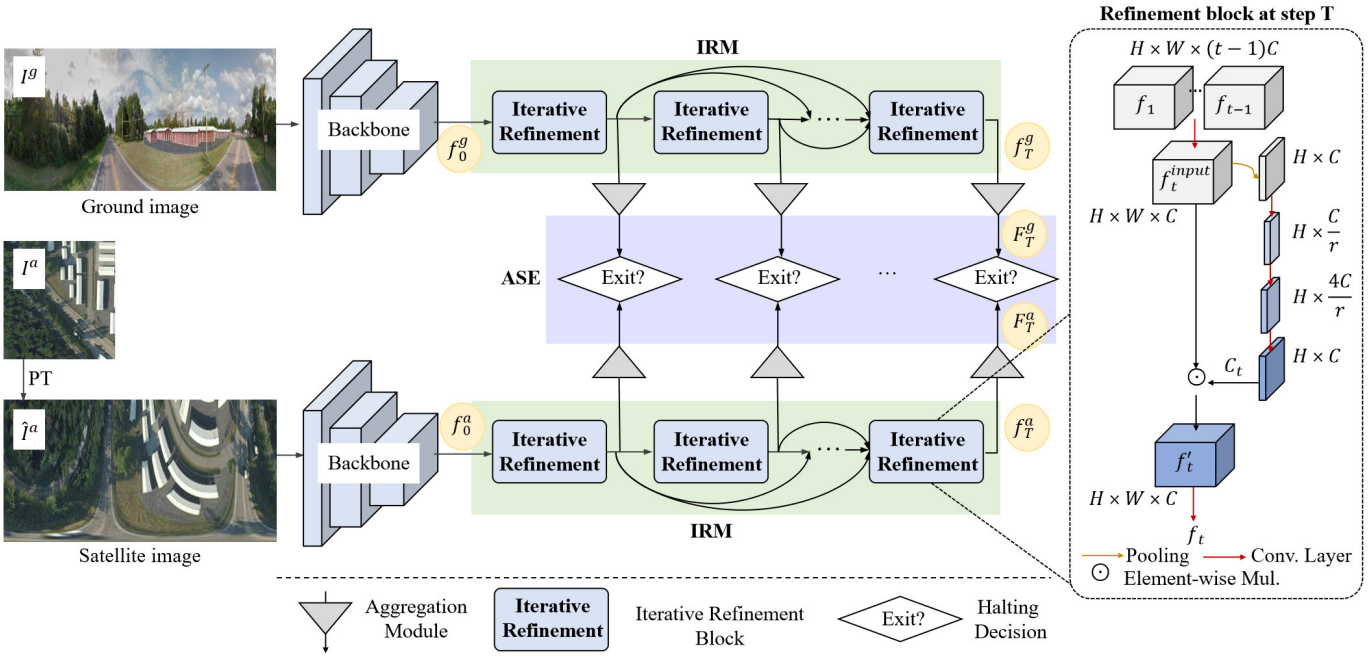

Fig. 4. Overview of our SIRNet, which consists of four components: the backbone networks, the IRMs (highlighted in green), the aggregation modules (gray triangle), and the ASE mechanisms (highlighted in purple). Specifically, the IRM contains T cascaded refinement blocks to improve the rough predictions output from the backbone step by step, and the structure of a single refinement block at step T is depicted on the right. The ASE, which is used solely at inference, enables the network to automatically adapt the number of refinement steps to each input query by making halting decision at each refinement step.

Some works have recently explored different variants of cross-view geo-localization tasks under drone-based setting [25], one-to-many setting [26], and orientation unaligned setting [11]. In this work, our proposed method focuses mainly on the standard setting, i.e., matching a ground panorama with a corresponding satellite image.

B. Iterative Refinement

The idea of iterative refinement is first proposed for human pose estimation [27], where body structure is modeled by learning a hierarchical feature encoder with a top-down iterative feedback mechanism. Iterative refinement mechanism has also been used for semantic segmentation [28], instance segmentation [29], image synthesis [30], [31], and object detection [32], [33]. To the best of our knowledge, we are the first to introduce an iterative refinement scheme for cross-view geo-localization. Moreover, our method differs notably from previous works in two aspects. First, most existing methods require pixel-level annotation to assist in refinement, which is costly and may not be available in practice. Instead, our approach requires only image-level supervision. Second, we are also the first to propose an ASE mechanism for such an iterative refinement framework. This makes our network a dynamic architecture and perform better than its static counterpart.

III. Methodology

In this section, we present the SIRNet, a novel step-adaptive iterative refinement method for cross-view geo-localization. We first give an overview of the proposed SIRNet in Fig. 4 and Section III-A. Then two key components of our model, i.e., the IRM and the ASE mechanism, are described in Sections III-B and III-C, respectively.

A. Network Overview

Let $I^g \in \mathbb{R}^{H^g \times W^g \times 3}$ and $I^a \in \mathbb{R}^{H^a \times W^a \times 3}$ denote a ground image and an aerial image, respectively, where H and W are the spatial dimensions of the image. In line with [5] and [6], we apply the polar transform, a simple trick that geometrically aligns ground and aerial images by warping the aerial images, as shown in Fig. 4. The warped aerial images are denoted as $\hat{I}^a \in \mathbb{R}^{H^g \times W^g \times 3}$. To extract ground and aerial representations separately, we adopt a Siamese-like network architecture with two independent branches of the same structure. For simplicity, we omit the superscript g or a in the later descriptions if not specified.

As shown in Fig. 4, the SIRNet consists of four important components: a backbone network, an IRM, several aggregation modules, and an ASE mechanism. We use the first eight layers of VGG16 [34] as the backbone network to extract an intermediate feature map $f_0 \in \mathbb{R}^{H_{\text{in}} \times W_{\text{in}} \times C_{\text{in}}}$, where H_{in}, W_{in}, and C_{in} are the height, width, and the number of feature channels, respectively. To generate more discriminative representations, we incorporate the IRM (described in Section III-B) at the end of the backbone network. The IRM regards the intermediate feature map f_0 as a rough network prediction and refines it progressively. Following that, we attach feature aggregation modules sharing their weights with all the refinement blocks (the structure of the aggregation module can be found in Section III-B). This allows the network to produce image representations at any refinement step, thus facilitating the ASE at inference. At inference, the ASE method adaptively

determines the number of refinement steps for each input sample by making termination decisions at every refinement step (elaborated in Section III-C).

B. Iterative Refinement Module

The ultimate target of the SIRNet is to generate discriminative ground and aerial representations. However, viewpoint variations and visual interferences pose great challenges. Motivated by the human perception process, as mentioned before, we propose an IRM to address these challenges by endowing the network with the capability of progressive self-correcting. In the following, we first illustrate the overall iterative refinement framework. Then, we elaborate on the internal structure of each refinement block.

1) Iterative Refinement: As shown in Fig. 4, the proposed IRM embodies the idea of progressive self-correcting by cascading T refinement blocks, which share the same structure without weight sharing. For each refinement block, the goal is to improve the output of the previous block, such that the IRM is capable of improving the rough network prediction, i.e., f_0, step by step (i.e., block by block). One natural choice for refinement is to supervise the IRM by pixel-level annotations of discriminative regions. Yet, preparing pixel-level annotations is very expensive. Inspired by the fact that context plays a crucial role in resolving visual ambiguity or incompleteness in the human perception system [35], we sidestep this problem by self-correcting the rough prediction with the aid of region-level scene context.

2) Single Refinement Block: Formally, given the input of the tth refinement block, i.e., $f_t^{\text{input}} \in \mathbb{R}^{H_{\text{in}} \times W_{\text{in}} \times C_{\text{in}}}$ ($t \in \{1, \ldots, T\}$), the refinement block firstly aggregates context features [see (1)] to assist in self-correcting. Afterward, the refinement map is generated by reconsidering which channels are critical in the region based on its context [see (2)]. Following that, we acquire the refined feature at step t by the element-wise multiplication of refinement map and f_t^{input} [see (3)]. Fig. 4 depicts this procedure. For a better understanding, we formulate this process as follows:

$$\mathcal{C}_t = \text{pool}_h\left(f_t^{\text{input}}\right) \tag{1}$$

$$\mathcal{A}_t = \text{Dup}\left(\text{conv}_{1\times1}^t\left(\text{conv}_{3\times3}^t\left(\text{conv}_{3\times3}^t(\mathcal{C}_t)\right)\right)\right) \tag{2}$$

$$f_t = \text{Drop}\left(\text{conv}_{3\times3}^t\left(\mathcal{A}_t \odot f_t^{\text{input}}\right)\right). \tag{3}$$

In (1), pool_h denotes an average pooling operation with $1 \times W_{\text{in}}$ pooling kernel. We term $\mathcal{C}_t \in \mathbb{R}^{H_{\text{in}} \times 1 \times C_{\text{in}}}$ as a height-wise context because each value of $\mathcal{C}_t$ represents the context of a horizontally divided region. Due to the structural nature of ground images (see Fig. 5), where each row of a scene image has a notably different object distribution [36], the height-wise context is more informative and discriminative for self-correction. In the experiment, we further discuss and compare three kinds of context, i.e., height-wise, width-wise, and local-wise contextual information. In (2), $\text{conv}_{k\times k}^t$ denotes a convolutional layer with kernel size $k \times k$ followed by a non-linear activation function, i.e., ReLU, at refinement step t. Dup indicates the duplicate operation used to ensure that the shape of $\mathcal{A}_t$ is the same as that of f_t^{input}. In this way, the generated $\mathcal{A}_t$ can be regarded as a refinement map for self-correcting. In (3), $\odot$ denotes the Hadamard product, and Drop indicates the dropout operation [37].

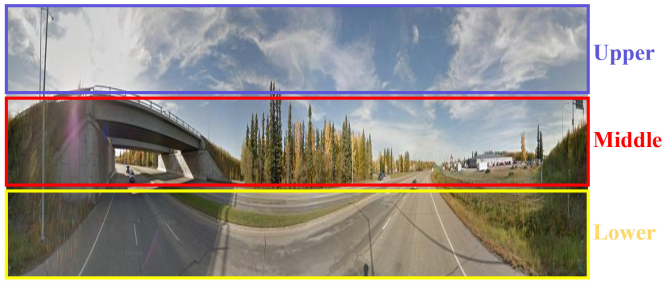

Fig. 5. Each part of an image divided into three horizontal sections has a significantly different object distribution from each other. For example, roads lie mainly in the lower region.

3) Aggregation Module and Dense Connectivity: At the end of each refinement block, we attach an aggregation module, as shown in Fig. 4. In line with [6], each aggregation module consists of three convolutional layers, reducing the height and channels of the feature maps but maintaining their width. By flattening the aggregated feature map, we acquire an image descriptor at step t, i.e., F_t, with the dimension of $4 \times 16 \times 64 = 4096$. It is worth mentioning that attaching the aggregation modules to all the refinement blocks enables the network to produce image representations at any refinement step, thus facilitating the ASE (elaborated in Section III-C). Nevertheless, as pointed out in [38], the introduction of aggregation modules at early refinement steps can harm the final aggregation module. Therefore, following the suggestion in [38], we densely connect each refinement module with all subsequent refinement modules to mitigate this problem. As shown in Fig. 4, for $t > 2$, the input f_t^{input} of the tth refinement block is gained by densely connecting the outputs of all subsequent refinement blocks using concatenation and convolution operations, and for $t \leq 2$, $f_t^{\text{input}} = f_{t-1}$.

C. Adaptive Step Estimation

In such an iterative refinement framework, a natural question is: how do we determine the refinement steps? Naturally, one can statically set the number of refinement steps to 3 or 4 [30], [39], [40]. However, we observe that setting a fixed number of refinement steps limit the model's performance because input images have different characteristics. Therefore, we propose an ASE algorithm that allows the network to adapt the number of refinement steps to each input.

1) ASE Algorithm: Instead of setting a fixed number of refinement steps, as shown in Fig. 4, we set the maximum number of refinement steps (i.e., the number of refinement blocks T). The ASE is performed only at inference by making a halting decision at every refinement step until the refinement is terminated or the maximum number of refinement steps is reached. Specifically, at each step, the SIRNet estimates whether the next refinement step brings improvement by comparing representations output from adjacent refinement steps (short for adjacent representations in the following) in terms of their softmax confidences. If the refined features

are discriminative enough for the task, the network outputs an image representation at the current step. Otherwise, the network continues with the next refinement step until the maximum number of refinement steps is reached.

Algorithm 1 Adaptive Step Estimation

Input: Intermediate features of a ground query f_0^g;
Database $D = \{(F_1^{a_1}, \ldots, F_1^{a_N}), \ldots, (F_T^{a_1}, \ldots, F_T^{a_N})\}$;
Maximum number of refinement steps T;
Output: Ground image descriptor F^g;

for $t \leftarrow 1$ *to* $T-1$ **do**
  // Extract adjacent image descriptors
  Get $\{F_t^g, F_{t+1}^g\}$;
  // Compute L2 distances
  $D_t \leftarrow [d(F_t^g, F_t^{a_1}), \ldots, d(F_t^g, F_t^{a_N})]$;
  $D_{t+1} \leftarrow [d(F_{t+1}^g, F_{t+1}^{a_1}), \ldots, d(F_{t+1}^g, F_{t+1}^{a_N})]$;
  // Compute softmax confidences
  $\{C_t\} \leftarrow -softmax(nsmallest(2, D_t))[0]$;
  $\{C_{t+1}\} \leftarrow -softmax(nsmallest(2, D_{t+1}))[0]$;
  if $C_t > C_{t+1}$ **then**
    // Terminate
    return F_t^g;
  end
end
return F_T^g;

Algorithm 1 shows the detailed ASE procedure as pseudo code. First, starting from the first refinement stage, i.e., $t = 1$, we extract the image descriptors $\{F_t^g, F_{t+1}^g\}$ output from the aggregation module of the current stage t and the next stage $t+1$, respectively. Second, we calculate the $L2$ distances between the ground and aerial database features output at stage t and stage $t+1$, respectively. Note that in the third and fourth lines of Algorithm 1, $F^{a_i}j$ denotes the descriptor of the ith aerial database image output at the jth stage. Third, we use the minimum value of the negative softmax top-2 distances as the confidence measure and calculate the confidence scores of F_t^g and F_{t+1}^g. Features with higher confidence scores are considered more discriminative. Finally, to make a halting decision, we estimate whether the next refinement step brings improvement. This is achieved by comparing the descriptor confidences between every two refinement steps. If the confidence score decreases after the next refinement step, we output the final descriptor at step t. Otherwise, we repeat the above process as t increases until $t = T-1$.

2) Multiscale Feature Augmentation: A significant challenge of the ASE is how to measure and compare the discriminative ability of representations. As described above, we adopt the negative softmax as a metric. In this section, we further introduce a multiscale feature augmentation technique that can significantly improve the accuracy of step estimation. The motivation is to take into account the information at different scales when comparing adjacent representations since the scale is a crucial factor affecting feature representation [41], [42]. This enables the ASE to comprehensively compare adjacent representations at different abstract levels, leading to better performance of step estimation.

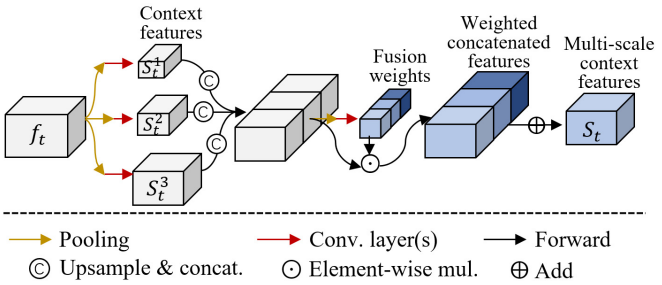

Fig. 6. Extracting multiscale context feature S_t for multiscale feature augmentation.

As shown in Fig. 6, we start by extracting several context features from the refined feature f_t at step t via the average pooling layer followed by a 1×1 learnable convolution layer. Note that the pooling layer has a kernel of shape $(H_{\text{in}} - s(\alpha H_{\text{in}} - 1)) \times (W_{\text{in}} - s(\alpha W_{\text{in}} - 1))$ and stride $s = \lfloor 1/\alpha \rfloor$. As a result, the output is a context feature of shape $\alpha(H_{\text{in}} \times W_{\text{in}})$, where α is a scale factor ranging between 0 and 1. By setting three different values of α, we obtain three context features (S_t^1, S_t^2, S_t^3) at different scales of $\alpha_k(H_{\text{in}} \times W_{\text{in}})$, where $k \in \{1, 2, 3\}$. Then, we aim to fuse (S_t^1, S_t^2, S_t^3) into a single multiscale context feature map. To do so, we upsample S_t^1, S_t^2 and S_t^3 to the same shape of $H_{\text{in}} \times (W_{\text{in}}/2)$ and concatenate them along the channel. Afterward, a global average pooling operation is performed on the concatenated features with three convolutional layers to learn the fusion weights. By multiplying the weights and concatenated features element-wisely and adding the weighted concatenated features along the channel, we can acquire the multiscale context feature S_t. Finally, we augment the output of the first convolutional layer in the aggregation module by S_t.

D. Training Objective

During the training phase, we apply a consistent weighted soft-margin loss, which enforces the same supervision signal on all refinement blocks. The overall loss function L consists of T terms, each of which indicates the weighted soft-margin loss [8] of a specific refinement step. Formally, given a triplet with a ground image I^g, its positive aerial exemplar I^a and a negative aerial exemplar I^{a*}, the loss function L can be computed as follows:

$$L = \sum_{t=1}^{T} L_t, \quad L_t = \log\left(1 + e^{\beta \cdot \left(d\left(F_t^g, F_t^a\right) - d\left(F_t^g, F_t^{a^*}\right)\right)}\right) \tag{4}$$

where the hyperparameter β is used to speed up training convergence, and d indicates the L_2 distance. F_t^g, F_t^a, and $F_t^{a^*}$ denote the image descriptors of ground image, the positive aerial image, and the negative aerial image output at stage t, respectively.

IV. Experiment

A. Dataset and Evaluation Protocol

1) Dataset: We evaluate our SIRNet on two widely used benchmark datasets, CVUSA [4] and CVACT [13]. CVUSA dataset consists of 35 532 image pairs for training and

8884 pairs for test. CVACT dataset provides the same number of image pairs for training, 8884 image pairs for validation, and 92 802 image pairs with accurate UTM coordinates (i.e., geo-tags) for testing. For clarity, we denote the CVACT validation and test sets as CVACT_val and CVACT_test, respectively. For CVUSA and CVACT_val, the correct match of a ground image is a single aerial image covering the same or wider region, while for CVACT_test, the correct matches of a query ground image are all aerial images within 5 m of the ground-truth location of the query image. In the following, we indicate the geo-localization tasks performed on CVUSA and CVACT_val as "standard" cross-view geo-localization and indicate the tasks performed on CVACT_test as "fine-grained" cross-view geo-localization.

2) Evaluation Protocol: To fairly compare with several state-of-the-art methods, we follow the evaluation protocol used in [5], [6], [13], and [15]. Specifically, we compute the recall accuracy at the top K ($r@K$ for short; $K \in \{1, 5, 10, 1\%\}$), which represents the probability of correct match(es) ranking within the first K results. For CVUSA, $r@1\%$ indicates the recall accuracy at the top 1% of the test set, and for CVACT_val and CVACT_test, $r@1\%$ indicates the recall accuracy at the top 1% of CVACT_val.

B. Implementation Details

Our SIRNet is implemented using TensorFlow [43]. During the training phase, we initialize the parameters of the backbone network with pretrained weights on ImageNet [44] and randomly initialize the remaining parameters. The overall network is trained end to end by applying the Adam optimizer [45] with a learning rate of $1e^{-5}$. The dropout rate is set to 0.8. If not specified, in the IRM, T is empirically set to 3 to achieve a better trade-off between model complexity and accuracy. In the multiscale feature augmentation module, the scaling factors α_1, α_2, and α_3 are set to 0.2, 0.3, and 0.4, respectively. For the loss function, β is set to 10. Batch size B is set to 32, and for each ground or aerial image in B positive pairs, there are $B-1$ negative pairs from all the other images, hence totally producing $2B(B-1)$ triplets.

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

1) Compared Methods: We compare the proposed SIRNet against 15 state-of-the-art methods on CVUSA [4], CVACT_val [13], and CVACT_test [13] datasets. For the compared methods, we directly cite the reported results from their articles. We choose the seminal works of Workman et al. [46], Vo and Hays [47], and Zhai et al. [4] that make the first effort to introduce CNNs to ground-to-aerial matching. We also compare our method with Siam-FCANet [7], Polar-SAFA [5] and the works of Liu and Li [13], Zhu et al. [11], Rodrigues and Tani [23] and Wang et al. [12], which learn discriminative representations via well-designed CNN-based models (e.g., incorporating attention mechanisms or orientation embeddings) or specialized metric learning techniques (e.g., loss function or data augmentation strategy). In addition, we select CVM-Net [8], the work of Regmi and Shah [16], CVFT [15], the work of Zheng et al. [25], Polar-SAFA [5], DSM [6], and the work of Toker et al. [17]. These works are designed to bridge the domain gap between the ground and aerial images, thus facilitating ground-to-aerial matching. It is worth mentioning that Polar-SAFA [5] and DSM [6] introduce a polar transform algorithm (a kind of data preprocessing technique) to align ground and aerial images in geometry coarsely. For fair comparisons with Polar-SAFA and DSM, as mentioned in Section III-A, we apply the polar transform to aerial images likewise. When comparing with works [4], [7], [8], [13], [15], [16], [25], [46], [47] that do not introduce the polar transform, we remove the polar transform from the SIRNet.

TABLE I

COMPARISONS WITH THE STATE-OF-THE-ART METHODS ON CVUSA [4] DATASET. THE BEST AND SECOND-BEST RESULTS ARE COLORED WITH RED AND BLUE, RESPECTIVELY. "PT" INDICATES WHETHER THE MODEL APPLIES (W/) THE POLAR TRANSFORM [5], [6] TO AERIAL IMAGES OR NOT (W/O). †: THE METHOD ADOPTS DATA AUGMENTATIONS DURING THE TRAINING PHASE

PT	Models	r@1 (%)	r@5 (%)	r@10 (%)	r@1% (%)
w/o	Workman [46] (ICCV15)	-	-	-	34.30
	Vo [47] (ECCV16)	-	-	-	63.70
	Zhai [4] (CVPR17)	-	-	-	43.20
	CVM-Net [8] (CVPR18)	22.47	49.98	63.18	93.62
	†Liu [13] (CVPR19)	40.79	66.82	76.36	96.12
	Regmi [16] (ICCV19)	48.75	-	81.27	95.98
	Siam-FCANet [7] (ICCV19)	-	-	-	98.30
	Zheng [25] (MM20)	43.91	66.38	74.58	91.78
	CVFT [15] (AAAI20)	61.43	84.69	90.49	99.02
	Zhu [11] (WACV21)	70.40	-	-	99.10
	†Rodrigues [23] (WACV21)	75.95	91.90	95.00	99.42
	Wang [12] (TCSVT22)	79.69	91.70	94.55	98.50
	our SIRNet	81.82	93.39	96.24	99.49
w/	Polar-SAFA [5] (NeurIPS19)	89.84	96.93	98.14	99.64
	DSM [6] (CVPR20)	91.93	97.50	98.54	99.67
	Toker [17] (CVPR21)	92.56	97.55	98.33	99.57
	our SIRNet	93.74	98.02	98.85	99.76

2) Standard Cross-View Geo-Localization: To test our proposed method on the standard cross-view geo-localization task, we compare the SIRNet with the state-of-the-art methods on CVUSA [4] and CVACT_val [13] datasets. We report the representative results ($r@1$, $r@5$, $r@10$ and $r@1\%$) in Tables I and II and present the complete $r@K$ curves in Fig. 7(a) and (b). The results show that the SIRNet achieves significantly higher performance than the compared methods. In particular, when applying the polar transform, the SIRNet gets the best performance of 93.74% $r@1$ on the CVUSA dataset and 86.02% $r@1$ on the CVACT_val dataset. When removing the polar transform, the SIRNet gains impressive $r@1$ performance of 81.82% on the CVUSA and 75.37% on the CVACT_val.

3) Fine-Grained Cross-View Geo-Localization: To examine the effectiveness of our method in the fine-grained setting, we compare our method against the advanced approaches on the CVACT_test dataset [13]. Since the images of CVACT_test densely cover a city with accurate GPS tags, learning discriminative representations to distinguish between similar-looking locations plays a crucial role in this task. As shown in Table III and Fig. 7(c), it is clear that the SIRNet outperforms the competing methods by a significant margin. Remarkably, the SIRNet exceeds the second-best method [17]

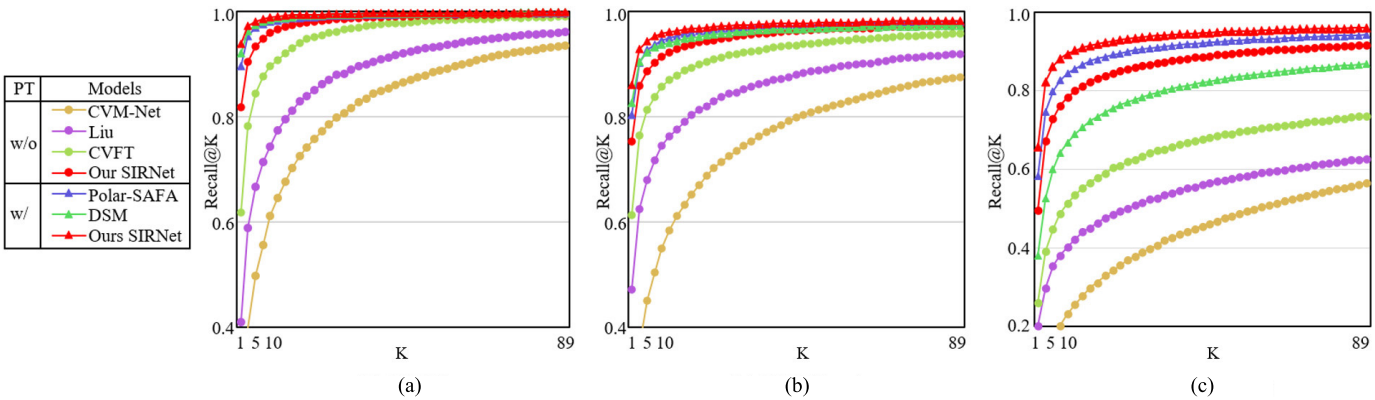

Fig. 7. Recall comparison at different values of K of our method versus the state-of-the-art methods on (a) CVUSA [4], (b) CVACT_val [13], and (c) CVACT_test [13] datasets. "PT" indicates whether the model applies (w/) polar transform [5] or not (w/o). Our models (marked in red) consistently outperform the competing methods across three datasets.

TABLE II

STANDARD CROSS-VIEW GEO-LOCALIZATION. COMPARISONS WITH STATE-OF-THE-ART MODELS ON CVACT_VAL [13] DATASET

PT	Models	r@1 (%)	r@5 (%)	r@10 (%)	r@1% (%)
w/o	CVM-Net [8]	20.15	45.00	56.87	87.57
	†Liu [13]	46.96	68.28	75.48	92.01
	CVFT [15]	61.05	81.33	86.52	95.93
	Wang [12]	73.85	87.54	90.66	95.87
	our SIRNet	75.37	88.76	91.90	97.42
w/	Polar-SAFA [5]	81.03	92.80	94.84	98.17
	DSM [6]	82.49	92.44	93.99	97.32
	Toker [17]	83.28	93.57	95.42	98.22
	our SIRNet	86.02	94.45	96.02	98.33

TABLE III

FINE-GRAINED CROSS-VIEW GEO-LOCALIZATION. COMPARISONS WITH STATE-OF-THE-ART MODELS ON CVACT_TEST [13] DATASET

PT	Models	r@1 (%)	r@5 (%)	r@10 (%)	r@1% (%)
w/o	CVM-Net [8]	4.06	16.89	24.66	56.38
	†Liu [13]	19.90	34.82	41.23	63.79
	CVFT [15]	26.04	44.91	52.25	73.59
	our SIRNet	49.66	72.84	78.35	91.63
w/	Polar-SAFA [5]	55.50	79.94	85.08	94.49
	DSM [6]	35.55	60.17	67.95	86.71
	Toker [17]	61.29	85.13	89.14	98.32
	our SIRNet	65.55	86.22	89.40	96.01

at $r@1$ by 4.26 points when applying the polar transform while outperforms the second-best method [15] at $r@1$ by a margin of 23.62 points without the polar transform. The results demonstrate the discriminative ability of the features learned by our models, highlighting the effectiveness of our method.

D. Ablation Study

1) Overall Ablation Study: To verify the effectiveness of each proposed component, we present overall ablation studies on the ASE and the IRM by gradually removing them from the SIRNet. For a fair comparison, all these variants adopt the same training setting as the SIRNet. As shown in Fig. 8 and Table IV, combining the network with the ASE and

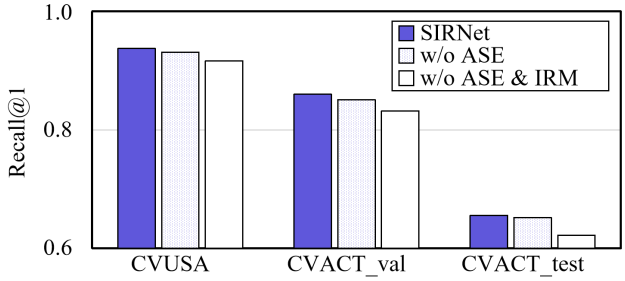

Fig. 8. Overall ablation studies on the IRM and the ASE on CVUSA, CVACT_val, and CVACT_test datasets.

TABLE IV

ABLATION STUDIES OF THE IRM AND THE ASE ON CVUSA DATASET

Models	r@1 (%)	r@5 (%)	r@10 (%)	r@1% (%)
SIRNet	93.74	98.02	98.85	99.76
w/o ASE	93.13	98.00	98.86	99.72
w/o ASE & IRM	91.67	97.24	98.21	99.63

the IRM consistently improves the performance across three datasets: CVUSA [4], CVACT_val [13], and CVACT_test [13]. Especially on the CVACT_val dataset, incorporating the IRM brings an improvement of 1.89%, while using the ASE improves the $r@1$ performance by a margin of 0.93%. The results demonstrate the effectiveness of our main contributions. In Section IV-F1, we provide additional qualitative results comparing our SIRNet with the baseline. In the following, we further conduct detailed ablation studies to analyze the impact of different designs for the IRM and the ASE.

2) Effect of the Maximum Number of Refinement Steps: In Table V, we study the impact of the maximum number of refinement steps by changing the numbers from 1 to 4. It is clear that the performance of the IRM is positively correlated with the maximum number of refinement steps when $T \leq 3$, while less of an improvement is noted when $T = 4$. It is straightforward that setting T to 3 provides the best trade-off between accuracy and complexity, allowing the network easier to train with fewer parameters. Therefore, we use $T = 3$ as the default choice for the SIRNet.

3) Comparing Three Kinds of Context: The contextual information used to assist in self-correcting is one of the key factors in the SIRNet. We also introduce and compare three

TABLE V

IMPACTS OF THE MAXIMUM NUMBER OF REFINEMENT STEPS IN THE IRM. THE "T" ITEM INDICATES THE MAXIMUM NUMBER OF REFINEMENT STEPS (I.E., THE NUMBER OF REFINEMENT BLOCKS)

T	r@1 (%)	r@5 (%)	r@10 (%)	r@1% (%)
1	93.20	98.00	98.80	99.70
2	93.42	98.05	98.83	99.67
3	93.74	98.02	98.85	99.76
4	93.63	98.27	98.80	99.74

TABLE VI

COMPARING THREE KINDS OF CONTEXT USED TO GUIDE THE ITERATIVE REFINEMENT. THE "CONTEXT" ITEM INDICATES WHICH CONTEXT IS USED TO ASSIST IN ITERATIVE REFINEMENT

Context	r@1 (%)	r@5 (%)	r@10 (%)	r@1% (%)
Width	93.14	97.93	98.67	99.66
Local	93.63	97.92	98.76	99.72
Height	93.74	98.02	98.85	99.76

kinds of contextual information, i.e., width-wise, local-wise, and height-wise context, used to guide refinement in the IRM. In Section III-B2, we introduce the height-wise context, and for comparison, here we extract width-wise context features by replacing the kernel in (1) with a $H_{\text{in}} \times 1$ average pooling kernel. In addition, Du et al. [48] have indicated that local features at neighboring spatial positions are semantic relevant as their receptive fields are highly overlapped. Inspired by this insight, we explore the local context for each pixel via the local-wise pooling kernel. Specifically, we replace the kernel in (1) with an average pooling kernel of shape $(H_{\text{in}} - s(0.3\ H_{\text{in}} - 1)) \times (W_{\text{in}} - s(0.3\ W_{\text{in}} - 1))$ and stride $s = 3$, yielding the context feature of shape $0.3(H_{\text{in}} \times W_{\text{in}}) \times C_{\text{in}}$. As shown in Table VI, we can find that three kinds of contextual information lead to better performance. Moreover, it is not surprising that due to the distinct structural priors of scene images, as mentioned in Section III-B2, the height-wise context-guided IRM outperforms the width-wise and local-wise context-guided IRMs.

4) Effect of the Dense Connectivity: As mentioned in Section III-B3, the dense connectivity of the IRM helps to weaken the negative impact of the aggregation modules at early refinement steps. To verify this point, we ablate the dense connectivity and report the ablation results in Table VII. We can see that with the effective bypassing paths, the dense connectivity can help increase the network's performance, which suggests the effectiveness of the dense connectivity in our network.

5) Effectiveness of the Multiscale Feature Augmentation: The multiscale feature augmentation is crucial for the network to make accurate step estimation since it enables the adjacent representations to be compared at different abstract levels. To illustrate this point, in Table VII, we ablate the multiscale feature augmentation. Results show that, by employing the multiscale feature augmentation, our network's performance is further boosted from 92.96% to 93.74% at $r@1$, demonstrating the effectiveness of the multiscale feature augmentation.

E. Effectiveness of Our Method

1) Incorporating With State-of-the-Art Methods: Fig. 9 demonstrates the effect of the step-adaptive iterative refinement method by incorporating our proposed modules (i.e., the IRM and the ASE) with three state-of-the-art models [5], [6], [15]. Note that we denote the original models as baseline models in Fig. 9 and insert our modules at the end of their backbone network. The modified networks are trained using the same settings as their baselines. From the results, we can clearly observe that the models incorporated with our proposed modules gain notable improvements, which demonstrates the effectiveness and wide applicability of our method. Significantly, combining the CVFT with our modules outperforms the baseline by +5.44%. It is also worth mentioning that although the Polar-SAFA introduces an attention mechanism to highlight salient features, combining the Polar-SAFA with our modules still improves the $r@1$ performance from 89.84% to 93.87%. This result implies that the single-step refinement method of Polar-SAFA may not be optimal for the ground-to-aerial matching task, and the proposed step-adaptive iterative refinement method is capable of enhancing the discriminative capability of the network.

TABLE VII

DETAILED ABLATION STUDIES OF THE DENSE CONNECTIVITY AND THE MULTISCALE FEATURE AUGMENTATION. THE "DC" AND "MA" ITEMS REPRESENT WHETHER WE INTRODUCE DENSE CONNECTIVITY AND MULTISCALE FEATURE AUGMENTATION INTO THE NETWORK, RESPECTIVELY

DC	MA	r@1 (%)	r@5 (%)	r@10 (%)	r@1% (%)
		92.49	97.57	98.60	99.65
✓		92.96	97.79	98.58	99.70
✓	✓	93.74	98.02	98.85	99.76

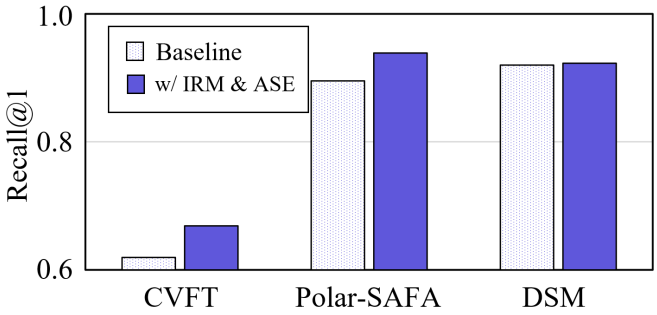

Fig. 9. Results of combining the proposed modules (the IRM and the ASE) with the state-of-the-art models.

2) Training With Fewer Samples: To further verify our method's generalization ability, we additionally evaluate its performance on few-shot cross-view geo-localization tasks. The goal of the few-shot task is to learn a model that can achieve generalization from only a small number of training examples [49]. This is more challenging than the standard cross-view geo-localization due to the unreliable empirical risk of a small number of samples. To support this task, we randomly select a certain percentage of samples from the CVUSA dataset [4] and generate four subsets accordingly. The size of each subset and its corresponding proportion to the CVUSA dataset are presented in the first two columns of Table VIII. For each dataset, we train the DSM [6] and our SIRNet from scratch and report recall accuracies tested on the original test set. As shown in Table VIII, our SIRNet consistently exceeds the DSM in the few-shot setting. In particular,

TABLE VIII

FEW-SHOT CROSS-VIEW GEO-LOCALIZATION ON CVUSA DATASET [4]. "# PAIRS" INDICATES THE NUMBER OF TRAINING IMAGE PAIRS SAMPLED FROM CVUSA, AND "PROP." INDICATES THE PROPORTION OF SAMPLING

# Pairs	Prop.	Models	r@1 (%)	r@5 (%)	r@10 (%)	r@1% (%)
7,106	20%	DSM [6]	80.15	91.76	94.57	98.84
		our SIRNet	84.89	94.34	96.24	99.28
14,212	40%	DSM [6]	86.98	94.99	96.87	99.36
		our SIRNet	89.18	95.97	97.59	99.59
21,319	60%	DSM [6]	90.16	96.39	97.73	99.46
		our SIRNet	91.69	97.19	98.24	99.62
35,532	100%	Polar-SAFA [5]	89.84	96.93	98.14	99.64
		DSM [6]	91.93	97.50	98.54	99.67
		our SIRNet	93.74	98.02	98.85	99.76

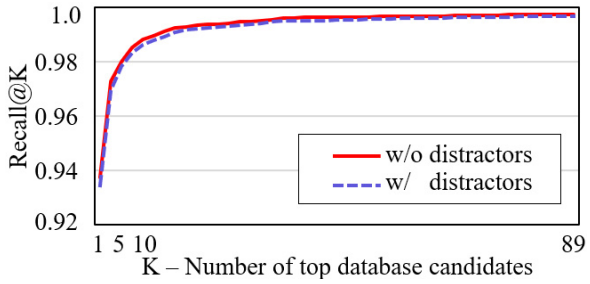

Fig. 10. Top-K recall accuracy on the test set with (w/) and without (w/o) distractor images. The model is trained on SIRNet on the CVUSA dataset.

the $r@1$ performance of the DSM drops 17.78 points when the number of training pairs decreases from 35 532 to 7106, while our SIRNet's performance at $r@1$ drops 8.85 points (51% less than that of the DSM). Furthermore, our SIRNet gains 89.18% at $r@1$ when training on 14 212 image pairs, which is competitive with the Polar-SAFA [5] gained by training on 35 532 pairs. The superior results indicate that our SIRNet can adapt well to the few-shot setting and is more widely applicable to real-world applications where data are sometimes scarce.

3) Adding Distractor Images: To evaluate whether our model is robust to distractor images, we add 8884 aerial images of the CVACT_val dataset [13] (disjoint with the CVUSA test set [4]) to the CVUSA test set and report the results of the SIRNet trained on the CVUSA dataset in Fig. 10. We find that even with distractor images, the SIRNet achieves impressive $r@1$ accuracy of 93.43% (a 0.31% performance drop). This experimental result demonstrates the stability and robustness of our proposed network in cross-view geo-localization task.

4) Comparing With the CBAM: From the perspective of reweighting features, our step-adaptive iterative refinement approach can be seen as a kind of attention mechanisms. To highlight the superiority of our method, we compare our proposed modules with a well-known attention-based module, i.e., CBAM [50], which learns spatial and channel attention masks to highlight salient features. Specifically, we replace the IRM and the ASE of our SIRNet with the CBAM as a competing model and train the competing model using the same settings as our SIRNet for a fair comparison. Then, we show their differences in Table IX. First, the key difference to CBAM lies in the step-adaptive iterative nature of our SIRNet. That is, the refinement process of our SIRNet is carried out in several refinement steps, and the number of refinement steps is specific to each input image. Such a framework endows the network with the capability of progressive self-correction, hence leading to better performance. Moreover, considering the particular structural priors of ground images, the SIRNet learns an attention mask specifically for each horizontally divided region. In contrast, the CBAM assigns different weights to each pixel. As a result, as shown in Table IX, the CBAM is inferior to the proposed modules on the CVUSA dataset [4] when the maximum refinement steps T is set to 1, 2, or 3, suggesting that the step-adaptive iterative refinement method is better suited for cross-view geo-localization task.

TABLE IX

COMPARISON BETWEEN THE CBAM AND OUR PROPOSED MODULES. T INDICATES THE MAXIMUM NUMBER OF REFINEMENT STEPS

Modules	Attention map	r@1 (%)		
CBAM [50]	spatial-specific	90.51		
IRM & ASE	spatially group-specific	$T=1$	$T=2$	$T=3$
		93.20	93.42	93.74

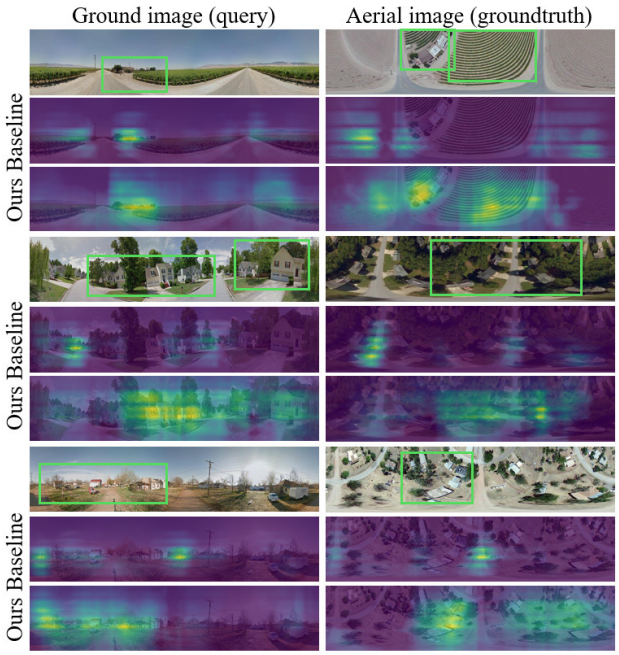

Fig. 11. Visualization of the generated features of our SIRNet and the baseline on the CVUSA [4] dataset.

F. Visualization Analysis

1) Comparing With Baseline: In Fig. 11, we compare our SIRNet with the baseline network on the CVUSA dataset by showing their generated heatmaps. Note that the IRM and the ASE are removed from the SIRNet as the baseline, and the baseline network is trained using the same settings as the SIRNet for a fair comparison. By default, we use Grad-CAM [51] for visualization. We show that our SIRNet is capable of activating the discriminative regions (e.g., houses) and depressing the less discriminative regions (e.g., roads) compared to the baseline network, which indicates the effectiveness of our method.

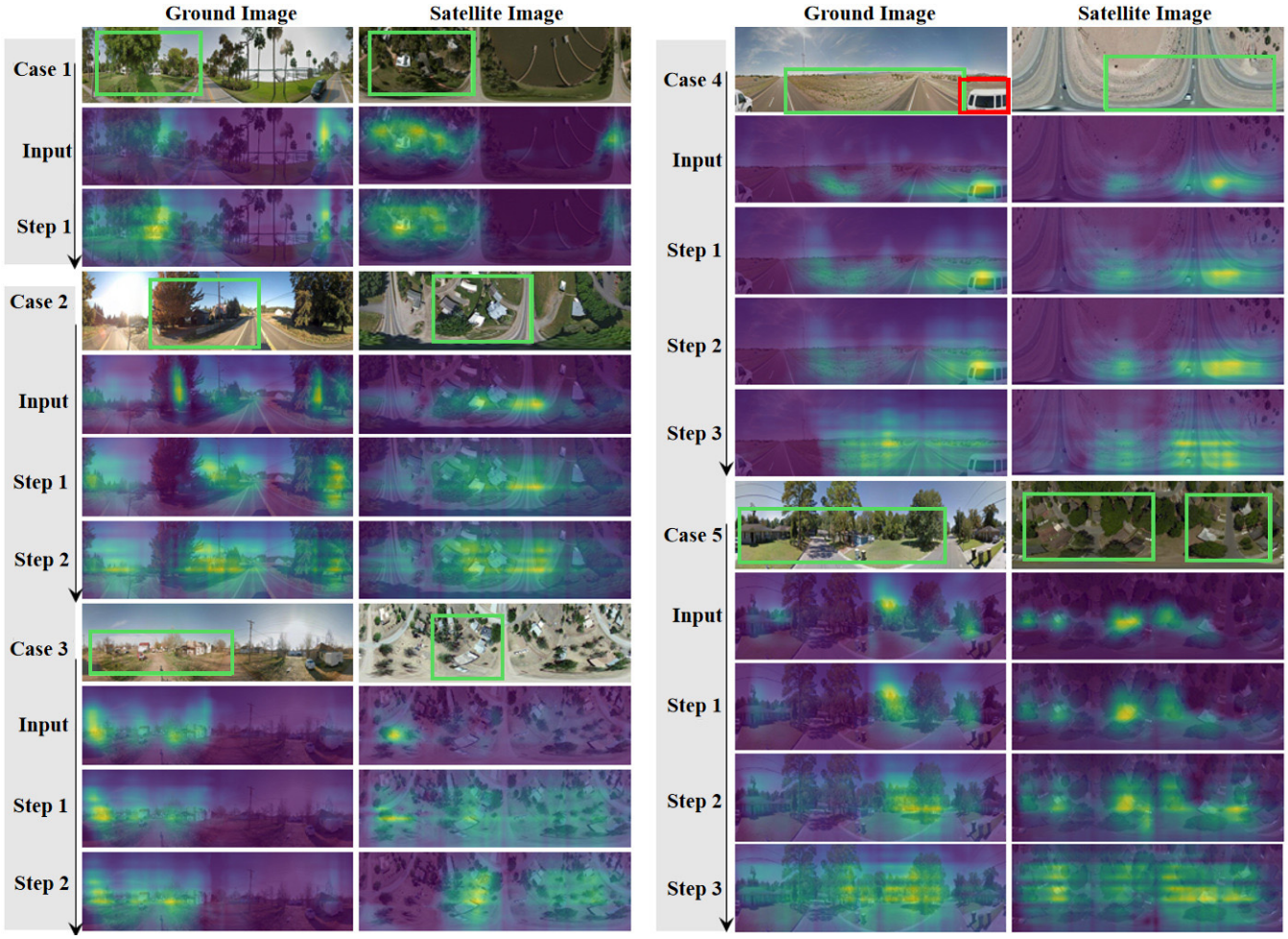

Fig. 12. Visualization of the generated features of the SIRNet on CVUSA [4] to verify the progressive self-correcting capability of the SIRNet. First, we visualize the rough network prediction (denoted as "Input") and the refined features of the IRM (denoted as "Step 1/2/3"). Second, we show that with the ASE, the SIRNet is capable of configuring the different number of refinement steps for each input sample (for cases 1, 2, and 3 and 4 and 5, the refinement steps are automatically set to 1, 2, and 3, respectively). Regions with higher activation values are indicated in yellow. For ease of reference, we box the discriminative regions and transient occlusions in green and red, respectively.

2) Verifying Our Motivation: We also visualize several refined feature maps of the IRM to verify whether the SIRNet has the capability of progressive self-correction. First, it can be seen in Fig. 12 that through step-adaptive iterative refinements, our SIRNet can capture more discriminative scene regions. Specifically, in Fig. 12 (case 1 and 2), our IRM is capable of highlighting houses even though they are partially shaded by trees, which enhances the discriminative ability of the learned features. In Fig. 12 (case 4), the original image feature map focuses mainly on a transient car that misleads cross-view matching at first, while our proposed IRM helps to suppress such interference. Second, in Fig. 12 (case 2) and Fig. 12 (case 5), we can also observe that the refinement process is progressive. That is, each stage assigns higher activation values to discriminative objects than its previous stage. Third, by presenting output feature maps at different stages, we also show that our SIRNet can estimate reasonable refinement steps via our proposed ASE. In summary, the above results highlight the progressive self-correcting capability of the proposed SIRNet.

3) Qualitative Results: Fig. 13 shows some qualitative ground-to-aerial geo-localization results using the best-performing model from Table III. For each ground query, we show its top-3 retrieved aerial images, which are accompanied with Universal Transverse Mercator Grid System (UTM) coordinates (i.e., the geo-tags). Images highlighted with green boxes are ground-truth retrieval results. We can see that our SIRNet is able to find aerial images that cover the same region

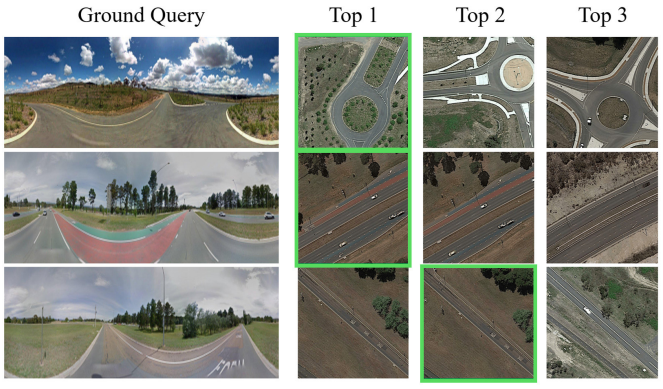

Fig. 13. Cross-view geo-localization results. These are top-3 retrieval aerial results of ground queries on CVACT_test dataset. The ground-truth images are in green boxes.

of the ground query images (e.g., the first and second rows), proving that the SIRNet indeed learns discriminative features.

V. Conclusion

In this article, we propose a novel SIRNet for the cross-view geo-localization task. First, the SIRNet includes an IRM, which can progressively refine the rough network predictions in several refinement steps. Second, we also propose an ASE mechanism, which automatically configures the number of refinement steps for each input sample. Experimental results show that our network outperforms the state-of-the-art methods. We also conduct extensive ablation studies on the proposed SIRNet to show its superiority. In addition, by incorporating our module with existing methods, training our model with fewer samples and adding distractor images at inference, we verify the wide applicability, generalization ability, and robustness of our method.

One main limitation of our method is that the ASE mechanism comes with additional computational overheads since it requires computing $L2$ distances when comparing adjacent representations. However, the computational costs can be greatly reduced by employing approximate nearest neighbor (ANN) [52] search instead of brute-force search. Since faster search is not the focus of this work, we leave this for future work.

References

[1] H. Bi, F. Xu, Z. Wei, Y. Xue, and Z. Xu, "An active deep learning approach for minimally supervised POLSAR image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9378–9395, Nov. 2019.

[2] Y. Zhu, J. Wang, L. Xie, and L. Zheng, "Attention-based pyramid aggregation network for visual place recognition," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 99–107.

[3] 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. (CVPR)*, Jun. 2016, pp. 5297–5307.

[4] 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. (CVPR)*, Jul. 2017, pp. 867–875.

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

[6] Y. Shi, X. Yu, D. Campbell, and H. Li, "Where am I looking at? Joint location and orientation estimation by cross-view matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4064–4072.

[7] 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. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8391–8400.

[8] S. Hu, M. Feng, R. M. H. Nguyen, and G. H. Lee, "CVM-Net: Cross-view matching network for image-based ground-to-aerial geo-localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7258–7267.

[9] B. Sun, C. Chen, Y. Zhu, and J. Jiang, "GEOCAPSNET: Ground to aerial view image geo-localization using capsule network," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2019, pp. 742–747.

[10] Y. Zhu, B. Sun, X. Lu, and S. Jia, "Geographic semantic network for cross-view image geo-localization," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.

[11] S. Zhu, T. Yang, and C. Chen, "Revisiting street-to-aerial view image geo-localization and orientation estimation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 756–765.

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

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

[14] S. Hu and G. H. Lee, "Image-based geo-localization using satellite imagery," *Int. J. Comput. Vis.*, vol. 128, no. 5, pp. 1205–1219, May 2020.

[15] Y. Shi, X. Yu, L. Liu, T. Zhang, and H. Li, "Optimal feature transport for cross-view image geo-localization," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 11990–11997.

[16] K. Regmi and M. Shah, "Bridging the domain gap for ground-to-aerial image matching," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 470–479.

[17] A. Toker, Q. Zhou, M. Maximov, and L. Leal-Taixe, "Coming down to Earth: Satellite-to-street view synthesis for geo-localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6488–6497.

[18] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.

[19] K. Regmi and A. Borji, "Cross-view image synthesis using conditional GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3501–3510.

[20] H. Tang, D. Xu, N. Sebe, Y. Wang, J. J. Corso, and Y. Yan, "Multi-channel attention selection GAN with cascaded semantic guidance for cross-view image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2417–2426.

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

[22] P. Li, P. Pan, P. Liu, M. Xu, and Y. Yang, "Hierarchical temporal modeling with mutual distance matching for video based person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 2, pp. 503–511, Feb. 2021.

[23] R. Rodrigues and M. Tani, "Are these from the same place? Seeing the unseen in cross-view image geo-localization," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 3753–3761.

[24] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 3859–3869.

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

[26] S. Zhu, T. Yang, and C. Chen, "VIGOR: Cross-view image geo-localization beyond one-to-one retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3640–3649.

[27] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik, "Human pose estimation with iterative error feedback," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4733–4742.

[28] C. Zhang, G. Lin, F. Liu, R. Yao, and C. Shen, "CANet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5217–5226.

[29] K. Li, B. Hariharan, and J. Malik, "Iterative instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3659–3667.

[30] C. Song, Z. Wu, Y. Zhou, M. Gong, and H. Huang, "ETNet: Error transition network for arbitrary style transfer," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 670–679.

[31] Y. Alaluf, O. Patashnik, and D. Cohen-Or, "ReStyle: A residual-based StyleGAN encoder via iterative refinement," 2021, *arXiv:2104.02699*.

[32] R. N. Rajaram, E. Ohn-Bar, and M. M. Trivedi, "RefineNet: Iterative refinement for accurate object localization," in *Proc. IEEE 19th Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2016, pp. 1528–1533.

[33] K.-W. Cheng, Y.-T. Chen, and W.-H. Fang, "Improved object detection with iterative localization refinement in convolutional neural networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2261–2275, Sep. 2018.

[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[35] I. Biederman, R. J. Mezzanotte, and J. C. Rabinowitz, "Scene perception: Detecting and judging objects undergoing relational violations," *Cogn. Psychol.*, vol. 14, no. 2, pp. 143–177, 1982.

[36] S. Choi, J. T. Kim, and J. Choo, "Cars can't fly up in the sky: Improving urban-scene segmentation via height-driven attention networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 9373–9383.

[37] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," 2012, *arXiv:1207.0580*.

[38] G. Huang, D. Chen, T. Li, F. Wu, L. van der Maaten, and K. Weinberger, "Multi-scale dense networks for resource efficient image classification," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–14.
[39] B. Cheng et al., "SPGNet: Semantic prediction guidance for scene parsing," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5218–5228.
[40] S. W. Zamir et al., "Multi-stage progressive image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14821–14831.
[41] W. Jiang, K. Huang, J. Geng, and X. Deng, "Multi-scale metric learning for few-shot learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 3, pp. 1091–1102, Mar. 2021.
[42] H. Wang, X. Hu, X. Zhao, and Y. Zhang, "Wide weighted attention multi-scale network for accurate MR image super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 962–975, Mar. 2022.
[43] M. Abadi et al., "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," 2016, *arXiv:1603.04467*.
[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.*, Jun. 2009, pp. 248–255.
[45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[46] S. Workman, R. Souvenir, and N. Jacobs, "Wide-area image geolocalization with aerial reference imagery," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3961–3969.
[47] N. N. Vo and J. Hays, "Localizing and orienting street views using overhead imagery," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 494–509.
[48] Y. Du, C. Yuan, B. Li, L. Zhao, Y. Li, and W. Hu, "Interaction-aware spatio-temporal pyramid attention networks for action classification," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 373–389.
[49] J. He, R. Hong, X. Liu, M. Xu, Z.-J. Zha, and M. Wang, "Memory-augmented relation network for few-shot learning," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1236–1244.
[50] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.
[51] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.
[52] W. Li et al., "Approximate nearest neighbor search on high dimensional data—Experiments, analyses, and improvement," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 8, pp. 1475–1488, Aug. 2020.

Xiufan Lu received the bachelor's and M.Sc. degrees from Shenzhen University, Shenzhen, China, in 2019 and 2022, respectively.

Her research interests include computer vision, image retrieval, and remote sensing image understanding.

Siqi Luo received the B.Sc. degree from Henan University, Kaifeng, China, in 2019, and the M.Sc. degree from Shenzhen University, Shenzhen, China, in 2022.

His research interests include computer vision, image retrieval, and remote sensing image understanding, especially on cross-view geo-localization.

Yingying Zhu (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees from Wuhan University, Wuhan, China, in 1998, 2001, and 2004, respectively.

She is currently a Professor with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. Her research interests include computer vision, remote sensing image understanding, geospatial data mining, and machine learning.