Learning to Find the Optimal Correspondence Between SAR and Optical Image Patches

Haoyuan Li, Fang Xu, Wen Yang, *Senior Member, IEEE*, Huai Yu, *Member, IEEE*, Yuming Xiang, *Member, IEEE*, Haijian Zhang, *Senior Member, IEEE*, and Gui-Song Xia, *Senior Member, IEEE*

***Abstract*—This study addresses the problem of finding the optimal correspondence for a given synthetic aperture radar (SAR) image patch from a large collection of optical reference patches, which is crucial for various applications, including remote sensing, place recognition, and aircraft navigation. However, achieving one-to-one SAR-Optical patch correspondence is challenging due to the distinct modal discrepancy and the poor discriminability of the target instances. To address these challenges, we propose a cross-modal patch correspondence scheme that consists of two modules: A retrieval-based coarse search module and a correspondence refinement module. Specifically, to explicitly represent the modal discrepancy, we first introduce a cross-modal adversarial learning strategy in the coarse search module and learn the modal-invariant feature embedding for retrieval. Furthermore, to improve the instance discriminability of retrieved candidates, we propose a graph representation in the refinement module to integrate the visual and spatial information, which is finally fed to an attention graph network to estimate the optimal correspondence. To evaluate the effectiveness of the proposed scheme, we also propose three new SAR-Optical patch correspondence datasets. Comprehensive experiments show that our approach significantly outperforms the competitors on all three datasets.**

***Index Terms*—Adversarial training, cross-modal image retrieval, graph neural network, synthetic aperture radar (SAR).**

I. Introduction

VISUAL localization is an important application of remote sensing [1], place recognition [2], and aircraft navigation [3], which is achieved by estimating the correspondence of the query and reference database images. This task is typically addressed as an image retrieval problem based on visual similarity. While the existing image-based retrieval methods have shown promising performance in scenarios where images are captured using optical cameras, they heavily rely on the assumption that optical images can always reliably capture the necessary information. However, this assumption may not hold true in challenging conditions such as low-light environments or adverse weather conditions. Hence, it becomes imperative to explore alternative and more robust information sources that can facilitate stable image correspondence even in such challenging scenarios.

The advantage of offering stable imaging during both day and night allows synthetic aperture radar (SAR) robust to adapt to light changes and variable weather on remote sensing. At the same time, optical satellite images are still the most popular and accessible archive, which can serve as the reference database for localization. Taking advantage of the imaging stability of SAR and the accessibility of optical satellite image archives for visual localization makes SAR-Optical patch correspondence a great potential task. However, SAR-Optical patch correspondence remains an underexplored research area. Optical sensors capture images by detecting reflected sunlight, whereas SAR sensors produce images by detecting backscattered waves from multiple microwave signals. The two types of images differ in radiation, noise level, and imaging geometry, resulting in distinct modality discrepancies. However, images originating from the same target/scene captured by different sensors should inherently possess consistent semantic information, which can be extracted using deep networks as modal-invariant features. Although the existing methods [4], [5], [6] proposed for cross-modal retrieval have succeeded in identifying the category of the query images, they primarily focused on treating the retrieval problem as an image classification task, which cannot distinguish between different places within the same category to meet the requirement of the patch correspondence. Therefore, it is necessary to develop new methodologies that go beyond the existing retrieval task to fully unlock the potential of SAR-Optical patch correspondence.

Concretely, two challenges of this task are listed as follows.

1) *Poor instance discriminability:* The patch correspondence requires optimal matching between the query and only one target. However, similar visual features lack instance discriminability, posing the challenge of identifying the

Manuscript received 22 May 2023; revised 7 September 2023; accepted 9 October 2023. Date of publication 16 October 2023; date of current version 27 October 2023. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62271355 and in part by the NSFC Regional Innovation and Development Joint Fund under Grant U22A2010. *(Corresponding authors: Wen Yang; Huai Yu.)*

Haoyuan Li, Fang Xu, Wen Yang, Huai Yu, and Haijian Zhang are with the School of Electronic Information, Wuhan University, Wuhan 430072, China (e-mail: lihaoyuan@whu.edu.cn; xufang@whu.edu.cn; yangwen@whu.edu.cn; yuhuai@whu.edu.cn; haijian.zhang@whu.edu.cn).

Yuming Xiang is with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China, and also with the School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: z199208081010@163.com).

Gui-Song Xia is with the School of Computer Science and the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430072, China (e-mail: guisong.xia@whu.edu.cn).

The project is available at https://Collebt.github.io/CMPC.

Digital Object Identifier 10.1109/JSTARS.2023.3324768

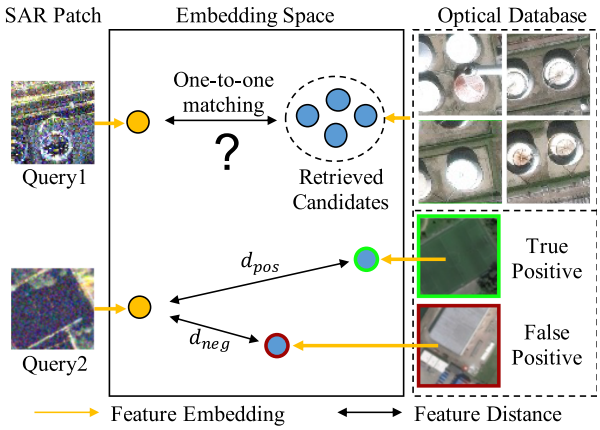

Fig. 1. Challenges of the CMPC task. Query1 shows challenge 1: The poor discriminability of the retrieved candidates on one-to-one correspondence. Query2 shows challenge 2: The modal discrepancy leads to the distinct feature distribution.

best correspondence from the retrieved candidates, as shown in query1 of Fig. 1. This limitation can lead to inaccurate and unreliable results in scenarios where the retrieved candidates have a high degree of similarity.

2) *Distinct feature distribution:* The fundamental differences in the imaging principles of SAR and optical modalities lead to variations in the appearance and structure of the same object in the two modalities, which makes it difficult to find a common representation for cross-modal features. As shown in query2 of Fig. 1, the negative sample is more similar than the true positive in embedding space due to the distinct feature distributions. It poses a challenge for learning-based methods that rely on the feature distribution metric. Overcoming this challenge requires the development of techniques that can efficiently model and bridge the feature distribution gap between these SAR and optical images.

Matching images for localization involves extracting feature descriptors from the images and computing similarity metrics to find the best correspondence. Over the last decades, feature description methods have been developed and proven helpful for retrieval between ground-to-satellite optical images [7], [8], [9], UAV-satellite optical images [10], [11], and cross-time place recognition [12], [13]. These methods are quite effective in matching query and reference images, which are both captured by optical sensors. However, their limitations become evident when facing inconsistent feature distribution across modalities, due to the large differences in imaging principles between SAR and optical images.

To address this challenge, several cross-modal retrieval methods [4], [14], [15], [16], [17] have been proposed. While these methods have shown remarkable performance in retrieving images across modality within scenes of specific categories, they retrieve multiple possible similar category candidates rather than a unique ground instance. It poses a challenge for applications that require precise localization and identification of specific objects. Since the GPS information is available in the reference archive, it provides an available way to refine the optimal correspondence by leveraging the location information of the reference candidates. This strategy might provide helpful spatial information to address the limitations of instance discriminability in category-level methods.

To overcome the aforementioned difficulties of SAR-Optical patch correspondence, we propose a coarse-to-fine correspondence scheme to explore the feasibility of instance-level cross-modal patch correspondence (CMPC). The proposed scheme comprises a cross-modal coarse search module and a refinement module. The coarse search module adopts adversarial learning to narrow the modal gap and extract modal-invariant features to retrieve the candidates. The refinement module turns the embedding features and the candidates' GPS information into a graph representation and then selects the optimal correspondence by updating the graph via an attention message propagation. To evaluate the performance of our proposed scheme, we also construct three SAR-Optical patch correspondence datasets.

In summary, our contributions are listed as follows.

1) We introduce a coarse-to-fine scheme for SAR-Optical remote sensing CMPC to find the optimal correspondence between SAR and optical images.
2) We explicitly model the cross-modal feature distribution as Wasserstein distance and propose a cross-modal adversarial learning strategy to learn the modal-invariant feature.
3) We propose a graph representation that incorporates the visual feature and spatial information to improve the discriminability of the retrieved candidates and refine the coarse retrieval to optimal correspondence.
4) We construct three datasets to evaluate various methods' feasibility of the CMPC task and even the localization applications. Our proposed scheme achieves state-of-the-art results on these proposed datasets.

The rest of this article is structured as follows. Section II offers a succinct survey of the related works. In Section III, we provide a comprehensive exposition of our proposed scheme, including the overview of the scheme, the cross-modal coarse search module, and the refinement module. Section IV presents and analyzes the experimental results. Section V discusses the limitations of the proposed scheme, as well as possible avenues for future research. Finally, Section VI concludes this article.

II. Related Work

In this section, we will review the recent progress in image-based retrieval, cross-modal category-level retrieval, and cross-modal instance-level retrieval.

A. Image-Based Retrieval

The task of image retrieval involves finding relevant images from a database of images given a query image [18]. This task has received significant attention in the research community in recent years. Since deep learning has been widely used to extract robust image features, Gong et al. [19] showed that the convolutional neural network (CNN) could effectively embed images into the global features for retrieval. Despite directly employing a developed model, Noh et al. [20] designed an

attention module on the vanilla model to strengthen the local features of the image. To leverage the advantage of both the local and global information, Song et al. [21] combined both local and global features to align different images and further improve retrieval effectiveness. To guide the network embedding the discriminative feature from images, a large number of metric learning methods have been proposed to regularize the distance between positive and negative samples. The core idea of these loss functions is to reduce the feature distance between the positive samples, as well as to enlarge the feature distance between the negative samples. Wen et al. [22] proposed the center loss, which distinguishes the feature center of each class of target during training. Schroff et al. [23] proposed a triplet loss to guide the network to learn an embedding distance between the positive samples and negative samples. Since the hard samples deteriorate the performance of the vanilla triplet, Hermans et al. [24] introduced the hard case mining strategy to make the triplet loss focus on the challenging samples. Sun et al. [25] unified the classification-based loss function and the distance-based loss function to improve retrieval effectiveness. However, these methods suffer from the domain gap when applied to the cross-modal retrieval task.

B. Cross-Modal Retrieval

Cross-modal image retrieval refers to measuring the similarity between images involving more than one modality. Due to the large visual appearance changes of the images from different types of sensors, the hand-crafted feature descriptor methods [26], [27], [28] encountered a bottleneck in the development of cross-modal image retrieval. Benefiting from the development of deep learning, recent works [29], [30] focus on learning modal-invariant features for both query and reference images from different modalities to improve the matching performance. To leverage the modal gap, Khokhlova et al. [31] adopted a Siamese network to extract modal-invariant descriptors of the multimodal images. In addition to extracting the modal shared feature, Liu et al. [13] proposed a separation network to extract modal exclusive features of the images from different domains. Ye et al. [32] employed a channel exchange strategy to switch the RGB image to a single-channel infrared image to reduce the color discrepancy between of two modalities. Jing et al. [12] improved a cross-modal center loss via a multilayer perception (MLP) to map different modality features into the mutual metric space. Huang et al. [33] considered that the positional relationships of the region are stable across different modalities and aligned the positional feature to improve the cross-modal matching accuracy. Facing the remote sensing sources, Li et al. [14] first proposed the cross-modal remote sensing category-level image retrieval dataset and employed the CNN to classify the panchromatic and multispectral images. Hash network [4] first solved the SAR-Optical category-level retrieval task by transforming the paired image to train the embedding network. However, these works focus on employing feature representations for classification, which cannot discriminate across instances, and thus are not suitable for instance-level retrieval.

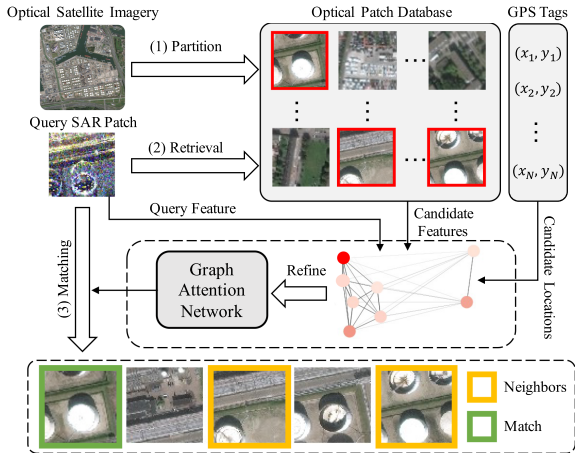

Fig. 2. Workflow of our proposed cross-modal remote sensing patch correspondence scheme.

C. Instance-Level Correspondence

Instance-level correspondence aims to identify the instances from images, such as localizing a street-view image to a satellite map. The main target of these works is to propose accuracy metric learning techniques to discriminate the instances. Several works [34], [35] focus on designing learning strategies for mining the discrimination between the instances. VIGOR [11] trained a network with the help of the neighbor patches of the target patches to estimate the image correspondence. In the situation when query images are from a new category, Yang et al. [36] showed that mapping the images into a uniform space would distort the manifolds of unseen classes, therefore designing a graph scheme to represent the feature space. With the help of Transformer [37], Tan et al. [38] mined the relationship between retrieved candidates by patch-based attention to rerank the retrieval results. In the SAR-Optical correspondence task, Hughes et al. [39] designed a pseudo-Siamese CNN to identify the established SAR-Optical patch correspondence. However, the discrepancy between SAR and optical images is too large to establish an optimal correspondence, leading to the unsatisfactory performance of the aforementioned methods.

III. Methodology

We propose a coarse-to-fine scheme to solve the task of SAR-Optical patch correspondence. The cross-modal coarse search module can be viewed as the cross-modal retrieval task, and the refinement can serve as the inlier estimation task. In this section, we present the overview of the proposed scheme, followed by a detailed description of the cross-modal coarse search module and the refinement module. Fig. 2 shows the overall flowchart of the proposed scheme.

A. Overview

In the retrieval step, deep cross-modal methods mainly reduce the impact of radiometric differences and speckle noise by dedicated network structures. However, designing a specific network

module would increase the model's complexity and reduce the model's robustness. Therefore, we propose to suppress these impacts in the training strategy without designing additional network modules.

First of all, we adopt random channel exchange transformation and image normalization as data augmentation. The random channel exchange can force the network to focus on contour and texture information shared by both optical and SAR images. Moreover, image normalization mitigates the impact of speckle noise by narrowing the dynamic range of the images, making it more visually interpretable and suitable for subsequent processing.

Second, we propose a cross-modal training strategy (see details in Section III-B) to guide the cross-modal embedding network f_{emb}, which maps the input images I_i to the modal-invariant feature x_i:

$$\boldsymbol{x}_i = f_{\text{emb}}(I_i). \tag{1}$$

After training the modal-invariant embedding network, the large optical satellite map is split into N patches and embedded to construct the reference vector set $\boldsymbol{X}^{\text{ref}} = \{\boldsymbol{x}_i^{\text{ref}}\}_{i=1}^{N}$ via f_{emb}. The SAR image I^{sar} is also embedded to a query feature $\boldsymbol{x}^{\text{que}}$ as well. The retrieval candidate set D is obtained by sorting the similarity in descending order, as shown in the following equation:

$$D = \{k | k \leq K_n, \boldsymbol{x}_k^{\text{ref}} \in \text{sort}(d(\boldsymbol{X}^{\text{ref}}, \boldsymbol{x}^{\text{que}}))\} \tag{2}$$

where K_n means the number of selected candidates.

Due to the poor instance discriminability between the query and reference images, the top retrieved candidate might not be the optimal correspondence for the query. Considering that the reference patches in the optical database typically contain the correct GPS location, we can leverage the location information of the reference patches to increase the discrimination of the retrieved candidates. Therefore, a refinement module (see details in Section III-C) is employed to address this issue and improve the initial retrieval results. Practically, we propose a graph representation G to incorporate the visual feature $\boldsymbol{x}$ and the location information $\boldsymbol{P}^{\text{ref}}$ of the retrieved candidates. The learnable refinement module f_{fine} finally estimates the inlier from this graph representation. The equation of refinement is shown as follows:

$$\hat{\boldsymbol{y}} = f_{\text{fine}}(G(\boldsymbol{x}^{\text{que}}, \boldsymbol{X}^{\text{ref}}, \boldsymbol{P}^{\text{ref}})) \tag{3}$$

where $\boldsymbol{P}^{\text{ref}} = \{\boldsymbol{p}_i^{\text{ref}}\}_{i=1}^{N}$ denotes the GPS coordinate of the reference patches. The final corresponding optical patch $I_{\text{match}}^{\text{opt}}$ can be selected from the highest predicted score of the refinement module, as shown in the following equation:

$$I_{\text{match}}^{\text{opt}} = \{I_i | \arg\max_i \hat{y}_i, i \in D\}. \tag{4}$$

B. Cross-Modal Coarse Search Module

To overcome the distinct modal discrepancy and extract modal-invariant features, we train the CNN with the Wasserstein adversarial learning strategy, combining it with the hard mining triplet and the feature projector, which aims to directly narrow the modal gap and learn the hard cross-modal samples. The coarse search module is shown in Fig. 3. The network's weights are shared between SAR and optical image embedding to enable the extraction of the mutual information from SAR and optical images.

1) Wasserstein Adversarial Training: In the cross-modal feature embedding, the significant disparity between two modalities results in differences between feature distributions, causing instability in cross-modal feature similarity measurement.

To address the challenge of the modal discrepancy in cross-modal feature representations, it is essential to model and reduce the gap explicitly. Besides employing the same shared network to extract the mutual information from SAR and optical images, we employ an adversarial discriminator to minimize the distance between extracted features from different modalities. The traditional classification discriminator only differentiates the modality to which the feature belongs, which does not measure the feature discrepancy. Instead, we introduce a Wasserstein discriminator to directly estimate the discrepancy between modalities. As cross-modal features belong to the distributions of their respective modalities, the Wasserstein distance can represent the discrepancy between the modal distributions by solving the earth-moving problem. Therefore, we employ the 1-D Wasserstein distance to explicitly model the cross-modal gap and introduce the Wasserstein adversarial learning to minimize the discrepancy between the modalities.

The 1-D Wasserstein distance between distributions $\mathcal{P}_s$ and $\mathcal{P}_t$ can be estimated by solving the optimal transportation problem. The definition of the Wasserstein discrepancy is given by

$$W_1(\mathcal{P}_s, \mathcal{P}_t) = \inf_{\gamma \in \Pi(\mathcal{P}_s, \mathcal{P}_t)} \mathbb{E}_{(x,y)\sim\gamma}[||x - y||] \tag{5}$$

where γ represents an optimal transportation from $\mathcal{P}_s$ to $\mathcal{P}_t$, and Π is the set of all couplings of $\mathcal{P}_s$ and $\mathcal{P}_t$.

In practice, we adopt the Kantorovich–Rubinstein duality [40] to approximate the original optimal transport problem (5), which avoids solving the bipartite matching problem iteratively

$$W_1(\mathcal{P}_s, \mathcal{P}_t) = \sup_{||f_w||_L \leq 1} \mathbb{E}_{x\sim\mathcal{P}_s}[f_w(x)] - \mathbb{E}_{x\sim\mathcal{P}_t}[f_w(y)]. \tag{6}$$

The equation demonstrates that maximizing the expectation of the optimal 1-Lipschitz function $||f_w||_L$ can approximate the Wasserstein discrepancy presented in (5). In the training phase, f_w acts as the Wasserstein discriminator, and we optimize it to achieve the maximum expectation of (6), which represents the 1-D Wasserstein distance. To train the Wasserstein discriminator, we first sample SAR and optical patches from the training batch as the two modal distributions and then use the negative Wasserstein distance as the loss function L_{dis}, as given in the following equation:

$$\begin{aligned} L_{\text{dis}}(\boldsymbol{x}) &= -W_1(\mathcal{X}_{\text{sar}}, \mathcal{X}_{\text{opt}}) \\ &= \sum_{\boldsymbol{x}_j \in \mathcal{X}_{\text{opt}}} f_w(\boldsymbol{x}_j) - \sum_{\boldsymbol{x}_i \in \mathcal{X}_{\text{sar}}} f_w(\boldsymbol{x}_i). \end{aligned} \tag{7}$$

During the training phase, the discriminator f_w maps the embedding features into the scalar space and calculates the expectation of the output scalars as the 1-D Wasserstein. The discriminator

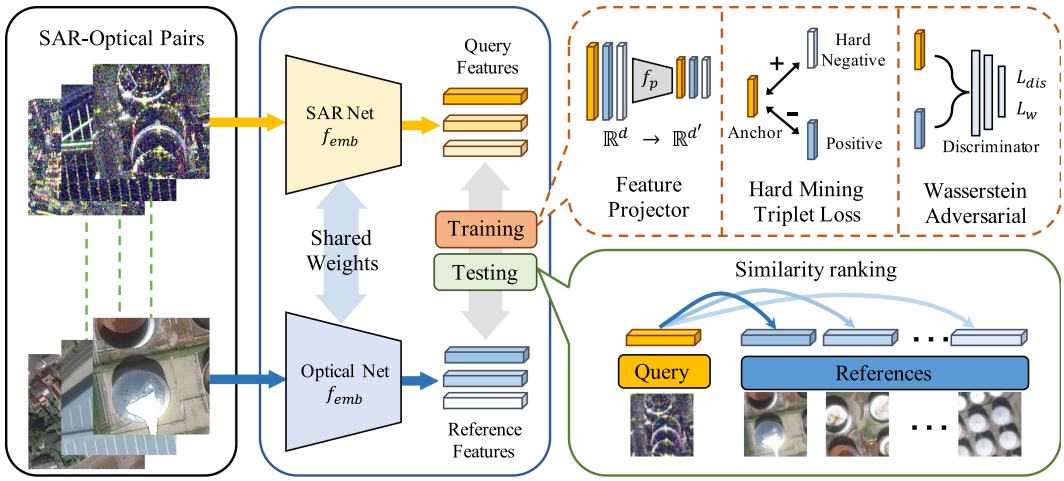

Fig. 3. Pipeline of the cross-modal coarse search module on the training and the testing phase. The training strategies are shown in the dashed red box. The testing phase is shown in the solid green box indicating coarse retrieval inference.

f_w also needs to fulfill the constraint of the 1-Lipschitz, which follows the regulation in WGAN [41]. It allows us to represent the discrepancy between different modalities through the output of the Wasserstein discriminator.

After T_d iterations of updating for the discriminator, L_{dis} can approach convergence, and its negative term can be used to approximate the Wasserstein discrepancy between the two modalities. Then, we can train the embedding network to minimize the discriminator output, effectively reducing the modality gap. The loss function for the feature embedding network is shown as follows:

$$L_w(\boldsymbol{x}) = W_1(\mathcal{X}_{\text{sar}}, \mathcal{X}_{\text{opt}}) = -L_{\text{dis}}(\boldsymbol{x}). \tag{8}$$

Finally, the embedding network is trained in an adversarial way. Specifically, we iterate the following two steps in the training phase: 1) estimate the Wasserstein discrepancy between output features of the embedding network by iteratively updating the discriminator f_w and b) update the embedding network f_{emb} by minimizing the estimated discrepancy $W_1(\mathcal{X}_{\text{sar}}, \mathcal{X}_{\text{opt}})$.

2) Hard Mining Triplet Loss: The distinct feature distribution across modalities can also lead to ambiguity between positive and negative samples. This issue may pose a challenge for the network to distinguish between the hard negative samples and positive samples. To address this issue, we employ the hard mining triplet loss, which selects only the hardest negative samples for each anchor sample. By focusing on the hardest negative sample, the model is forced to embed the discriminative features that can better distinguish between positive and negative samples in challenging cross-modal scenarios. In addition, the hard mining triplet loss reduces the computational complexity of the training process by eliminating common negative samples that are less informative.

Specifically, a training batch typically consists of B image pairs $\{I_i^{\text{sar}}, I_i^{\text{opt}}\}_{i=1}^{B}$. The embedded feature of each image in the batch is set as the anchor $\boldsymbol{x}_a$. Its matched patch in the other modality is used as the positive sample $\boldsymbol{x}_p$, while the nonmatched patches are negative samples. We select the most similar feature from negative samples as the hard negative sample $\boldsymbol{x}_n$. Therefore, we can sample $2B$ triplet sets consisting of three features $\{\boldsymbol{x}_a^i, \boldsymbol{x}_p^i, \boldsymbol{x}_n^i\}_{i=1}^{2B}$ that are used in the loss function, as shown in the following equation:

$$L_{\text{tri}} = \frac{1}{2B}\sum_{i=1}^{2B} \max(||\boldsymbol{x}_a^i - \boldsymbol{x}_p^i||_2^2 - ||\boldsymbol{x}_a^i - \boldsymbol{x}_n^i||_2^2 + \beta, 0) \tag{9}$$

where β means the margin distance. This loss function ensures that the selected negative samples are the most difficult to distinguish for the anchor within the batch, which can force the network to learn the feature that can better discriminate between the anchor and the challenging samples.

3) Feature Projector: The distinct feature discrepancy between SAR and optical makes the network lack of robustness and easily show overfitting. Therefore, a learnable feature projector f_p is employed in the training phase to improve the robustness of the network in cross-modal tasks. The projector f_p maps the output feature $\boldsymbol{x}$ to a normalized lower dimensional space with dimension d', as shown in the following:

$$f_p(\boldsymbol{x}) = \frac{\text{Norm}(\boldsymbol{W}\boldsymbol{x})}{||\text{Norm}(\boldsymbol{W}\boldsymbol{x})||_2}, \quad \boldsymbol{W} \in \mathbb{R}^{d\times d'} \tag{10}$$

where $\text{Norm}(\cdot)$ is the batch normalization utilized only on training phase. The batch normalization is only utilized in the training phase for regularization.

Our final loss function for the embedding network is shown as follows:

$$L_{\text{emb}} = L_{\text{tri}}(f_p(\boldsymbol{x})) + \lambda L_w(f_p(\boldsymbol{x})) \tag{11}$$

where the parameter λ controls the weight of the Wasserstein distance loss.

C. Correspondence Refinement Module

Benefiting from the modal-invariant feature embedding, the features can overcome the modal discrepancy and retrieve the candidates belonging to the same category. However, the visual

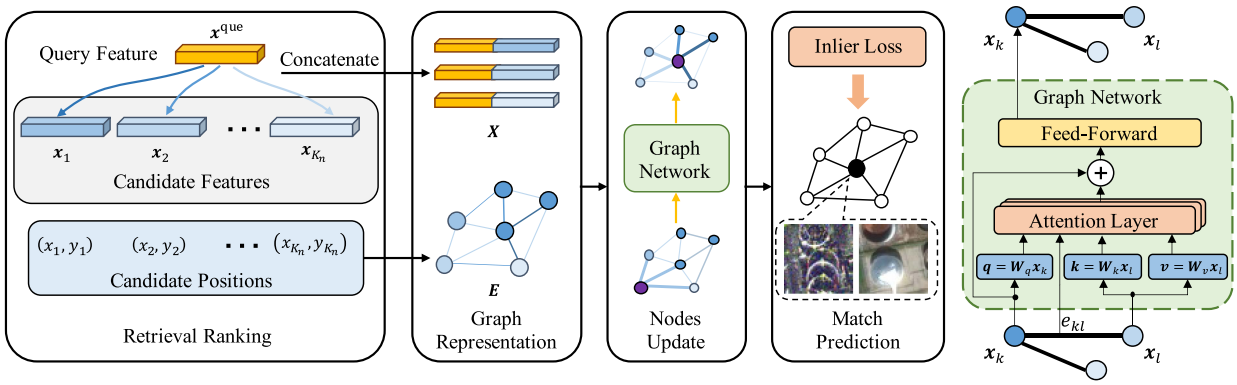

Fig. 4. Proposed refinement module. The graph representation combines the information of feature pairs and reference locations. The attention network updates the nodes to predict the inlier as the final optimal correspondence between the query and the references.

feature cannot be clearly corresponded to an instance within the same category, due to the poor discriminative of the visual and texture feature. Considering that the GPS tag of the reference patch is provided and represents the distinct location information of the retrieved candidates, we can mine the position relationship between the retrieved candidates to obtain the final match. We also concatenate the embedded features of both the query and the reference as node features to mine the mutual information.

Specifically, for each query image and its top-K_n matched candidates, we construct a graph with K_n nodes, where each node has K_e connections with its nearest neighbors based on the distance of the reference location. This graph structure allows us to perform message passing between nodes to refine their features and estimate the inlier probability of each node using a graph network. Fig. 4 shows the graph representation refinement module.

1) Graph Representation for Matching Pairs: The coarse search module produces a set of top-K_n feature candidates $\boldsymbol{x}_k \in \{\boldsymbol{x}^{\text{opt}}\}_{k=1}^{K_n}$, which are ranked by the similarity of the embedded features. However, there exists only one best matched reference patch from the retrieved candidates, so the module must predict the final matching probability of the candidate pairs. To do this, the information of the query feature and the reference features is combined as the graph node and fed into the network for prediction. Specifically, we concatenate the query feature $\boldsymbol{x}^{\text{que}}$ and the reference candidate feature $\boldsymbol{x}_k$ along the channel dimension into the node feature $\boldsymbol{x}_k^0$:

$$\boldsymbol{x}_k^0 = \sigma(C_1(\boldsymbol{x}^{\text{que}} || \boldsymbol{x}_k^{\text{opt}})), \quad k \in D \tag{12}$$

where $||$ denotes a concatenation operator, $C_1(\cdot)$ means 1×1 convolutional operation, and $\sigma(\cdot)$ means ReLU activation function.

After aggregating the matching pairs into input features for the refinement module, the node with the highest score represents the optimal correspondence from the retrieved candidates. Although the queries do not contain location information, the reference patch database can provide the correct GPS location. Therefore, we design the graph representation to take advantage of the reference GPS location with rich geometric information. We extract the position information from retrieved candidates' GPS tags and set it as the geometric position of the node feature. We calculate the distance between nodes and sort them in ascending order. The K_e nearest neighbors of each node can then be connected, and the feature of each edge is derived from the distance, as shown in the following equation:

$$e_{kl} = \begin{cases} \exp\left(\frac{-\|\boldsymbol{p}_k - \boldsymbol{p}_l\|_2^2}{\sigma_e}\right), & l \in \mathcal{N}(k) \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

where $\boldsymbol{p}$ is the GPS coordinate of retrieved candidate patches, $\mathcal{N}(k)$ are the neighbors of node k, defined as the K_e nodes with the closest geometric distance to k, and σ_e is the hyperparameter to control the scale of the connection weight.

After defining the node features and edge features, a graph representation can be built for the retrieval results, which transforms the refinement problem into an inlier selection problem.

2) Graph Attention Network: The ultimate goal of SAR-Optical patch correspondence is to find the best correspondence prediction from the reference patch for every query. After transforming the matching pairs into a graph representation with nodes $\boldsymbol{X}$ and edge connection $\boldsymbol{E}$, we propose a graph attention network f_{gnn} to estimate the inlier probability $\hat{\boldsymbol{y}}$ of each node, as shown in the following equation:

$$\hat{\boldsymbol{y}} = f_{\text{gnn}}(\boldsymbol{X}, \boldsymbol{E}). \tag{14}$$

To improve message propagation in the graph and leverage the effectiveness of the attention mechanism, we employ graph attention layers for message propagation, as shown in the right side of Fig. 4. Three learnable linear transformations first project the feature to the query $\boldsymbol{q}_k$ for node k, and key $\boldsymbol{k}_l$, value $\boldsymbol{v}_l$ for its neighbor l, respectively, as shown in the following equation:

$$\begin{aligned} \boldsymbol{q}_k &= \boldsymbol{W}_q \boldsymbol{x}_k \\ \boldsymbol{k}_l &= \boldsymbol{W}_k \boldsymbol{x}_l \\ \boldsymbol{v}_l &= \boldsymbol{W}_v \boldsymbol{x}_l. \end{aligned} \tag{15}$$

To update the kth node feature, we compute an edge attention weight α_{kl} for each of its connected neighbors $l \in \mathcal{N}(k)$ by taking into account the query vector $\boldsymbol{q}_k$, the key vector $\boldsymbol{k}_l$, and

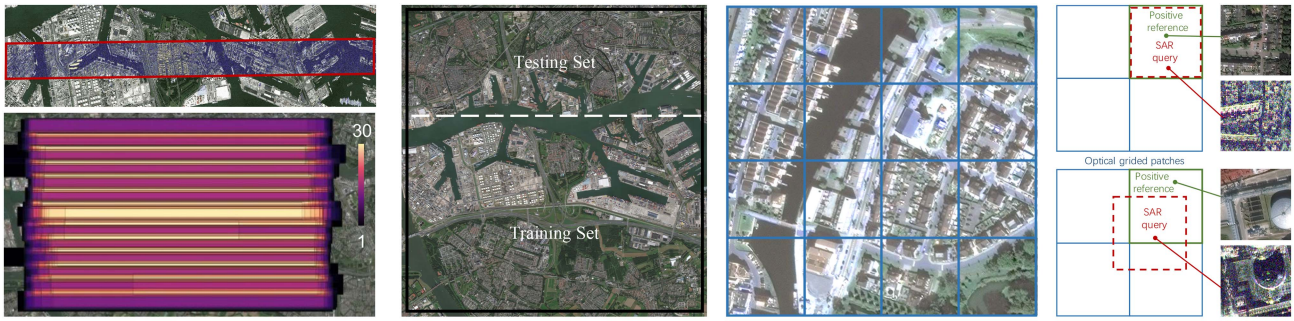

Fig. 5. Sampling strategy of the proposed datasets. (a) SAR and optical image data. (b) Nonoverlap region protocol for the training set and the test set. (c) Cropping strategy of optical patches. (d) Patch correspondence definitions.

the edge weight e_{kl}, as denoted in the following equation:

$$\alpha_{kl} = \text{softmax}(e_{kl}\boldsymbol{q}_k\boldsymbol{k}_l^\top). \tag{16}$$

This mechanism enables differential attention allocation to varying levels of information, allowing for the prioritization of important neighbor features. Then, we add the values $\boldsymbol{v}$ from the neighboring nodes to the original node feature and update it using a feedforward MLP f_θ:

$$\boldsymbol{x}_k^{t+1} = f_\theta\left(\boldsymbol{x}_k^t + \sum_{l\in\mathcal{N}(k)} \alpha_{kl}\boldsymbol{v}_l\right) \tag{17}$$

where $t = 0, 1, 2, ..., T$ is the step of update time; the node features update T times iteratively by catching the message from neighbors. The attention layers output the embedded feature $\boldsymbol{x}^T$ of every node. The final linear projection W_f maps the embedded feature to a scalar $\hat{y}$ as the inlier score of the node, as shown in the following equation:

$$\hat{y} = W_f\boldsymbol{x}^T. \tag{18}$$

The highest score from the nodes means the optimal correspondence between the query feature and the corresponding reference feature.

3) Inlier Loss: The refinement module aims to estimate the inlier node, which represents the optimal correspondence from the retrieved candidates. To accomplish this, we utilize the cross-entropy loss function L_{ce} to quantify the dissimilarity between the predicted probability $\hat{y}$ and the true inlier label y, as shown in the following equation:

$$L_{\text{ce}} = -\sum_{k=1}^{K_n}(y_k\log(\hat{y}_k) + (1-y_k)\log(1-\hat{y}_k)). \tag{19}$$

Specifically, the ground-truth label y_k is set to 1 if the retrieved candidate k is the true correspondence and 0 otherwise. By minimizing the loss function, the refinement module can effectively distinguish the inlier node from the outlier nodes and improve the accuracy of the retrieval results.

IV. Experiments

A. Experimental Setup

1) Patch Correspondence Datasets: Our proposed patch correspondence dataset is based on the SpaceNet 6 data [42], which consists of 204 different strips of SAR data collected from Rotterdam, The Netherlands, covering over 120 km^2 of the city. The SAR strips cover every area of the city, with at most 30 strips overlapping in some areas, as illustrated in Fig. 5(a). Each strip has four different polarization modes (HH, HV, VH, and VV), and we use three of these modes (HH, VV, and VH) to construct the polarized SAR pseudo-color images with a spatial resolution of 0.5 m. The optical imagery is acquired by the Maxar Worldview-2 satellite, with a spatial resolution of 0.5 m as well. We design the optical patch references and the SAR queries from these data and design three datasets to evaluate the proposed method. We manually align the SAR strips with the optical map to ensure that the cross-modal images are consistent. The GPS information of both the optical image and the SAR strips is manually corrected as well.

a) Optical patch archive: Given an optical map of the city, our objective is to match a SAR query in this map by searching the reference patches cropped from this map. To ensure that all queries can be matched with the reference patches, the reference patches must cover the entire map seamlessly. As shown in Fig. 5(c), the reference patches are cropped at a grid without any overlap. Every patch has a unique identification (ID) and has a fixed size of 200 × 200 pixels.

b) SAR patch queries: In real-world applications, the SAR strips captured by aircraft are not perfectly aligned with the optical map and have varying sizes. We design two types of SAR patches: aligned patches and nonaligned patches to construct the matching pairs. The IDs of SAR patches are the same as the matched optical patches.

1) *Aligned pairs:* The cropped SAR patches are aligned with the optical patches and have the same image size. The SAR patches that fall in the boundary of the strip are discarded due to the image being incomplete. The optical patches aligned with SAR patches are set as the positive samples of the SAR query, as shown at the top of Fig. 5(d).

TABLE I
DETAILS OF THE PROPOSED SAR-OPTICAL PATCH CORRESPONDENCE DATASETS

Dataset	Region overlap	Pair align	Image type	Train		Test	
				ID	Samples	ID	Samples
Aligned dataset	×	✓	SAR optical	2424 6328	18176 6328	873 2825	6487 2825
Same-area	✓	×	SAR optical	3355 9153	23080 9153	3027 9153	9240 9153
Cross-area	×	×	SAR optical	2475 6328	23080 6328	958 2825	9240 2825

2) *Nonaligned pairs:* All the SAR patches are cropped at a grid of size 200×200 pixels without being aligned to the optical map. These SAR patches are labeled with ground-truth GPS tags which are only used for correspondence identification. As shown at the bottom of Fig. 5(d), the green patch is considered as ground truth, which has the nearest GPS to the SAR query and contains the most shared objects with the query image.

c) Dataset protocol: We design two protocols for assigning the training set and the test set on the experiments: The overlap setting and the nonoverlap setting, according to different application scenarios.

1) *Region overlap:* All the optical patches are included for both the training phase and the testing phase. And the SAR patches are randomly split into two disjoint sets. This setting is for evaluating the methods when the data of the city is available in training.
2) *Region nonoverlap:* To assess the generalizability of the proposed scheme to new cities, we design the dataset protocol to ensure that the test region is not learned in the training phase. Fig. 5(b) shows that the patches are separated into two regions.

Above all, we design three datasets depending on whether the matching pairs are aligned and whether the testing region is available in training, as shown in Table I.

1) The *Aligned dataset* contains the paired SAR query aligned with the optical reference patches, which aims to evaluate the instance consistency of the cross-modal features.
2) The *Same-Area dataset* contains the nonaligned pairs, where all the optical patches are available in both the training and testing phases, which is focused on application scenarios when the city data is available for training.
3) The *Cross-Area dataset* represents a general challenge for methods to match pairs in a new region, which contains nonaligned pairs from no overlapping regions between the test set and the training set.

2) Evaluation Metrics

For the retrieval performance, we adopt the precision of top K in retrieval (P@K) and mean average precision (mAP) as the evaluation metrics. For geolocalization, we employ meter-level accuracy to evaluate the localization capability of nonaligned SAR patches. During the experiments, we evaluate the retrieval and geolocalization performance of the methods on three datasets. The best results in the experiments are bolded. All methods are trained using the training set and tested on the test set of each corresponding dataset.

1) *Top-K precision:* P@k computes the number of queries where the ground-truth label is among the top k label prediction. The definition is given by

$$\text{P@K} = \frac{\sum_{i=1}^{M} Acc_k}{M} \tag{20}$$

where Acc_k equals 1 if the ground-truth matched target is included in top k retrieval candidates, 0 otherwise. M is the number of all queries.

2) *mAP:* The average precision (AP) of each query i is the order in which the retrieved target is presented, and the mAP is the mean values for a retrieval result over a set of queries. The definitions are shown as follows:

$$\text{AP}_i = \sum_{k=1}^{N} \frac{\text{rel}(k)_i}{k}$$
$$\text{mAP} = \frac{1}{M} \sum_{i=1}^{M} \text{AP}_i \tag{21}$$

where $\text{rel}(k)_i$ is an indicator function equaling 1 if the item at rank k is a ground-truth target of query i, 0 otherwise. N is the patch number of the reference archive.

3) *Meter-level localization accuracy:* Localization accuracy evaluates the real-world distance between the predicted location and the ground-truth GPS location of the SAR query.

3) Implementation Details

The experiments are conducted on a platform equipped with 4x NVIDIA TITAN Xp GPUs. We adopt the stochastic gradient descent as the optimizer of the proposed scheme. All backbones on the experiments are pretrained with a batch size of 24 on the ImageNet dataset [43] and re-trained for 60 epochs. During the evaluation, we set the Top-1 matched target's geolocation as the localization prediction for the query patch.

We utilize the Res-Net50 as our feature embedding network on the coarse search module. β in L_{tri} is set to 0.3, and λ on L_{emb} is set to 0.01. The update time T_d for the discriminator is set to be 5, and the update time T of the refinement attention network is set to 3. We first train the coarse search module with L_{emb}. Then, we train the refinement module on the same training set with L_{ce}.

B. BENCHMARKING RESULTS

1) Comparison With Retrieval Methods

To evaluate the performance of our proposed CMPC, we compare it with several retrieval benchmarks on our proposed SAR-Optical patch correspondence datasets, including the geolocalization methods and the cross-modal methods.

TABLE II
METHOD COMPARISON ON THE ALIGNED DATASET

Model	P@1↑	P@5↑	P@10↑	P@20↑	mAP↑
Dual-constrained [44]	45.10	74.90	84.35	91.30	58.54
ReIDSB [45]	60.85	88.35	93.50	96.60	72.92
X-Modality [46]	67.45	91.45	95.30	96.85	77.80
AlignGAN [47]	69.30	92.10	96.20	97.15	79.48
CMPC (Ours)	**81.95**	**95.55**	**96.75**	**98.20**	**85.98**

The bold values indicate the top performing entries.

TABLE III
METHOD COMPARISON ON THE SAME-AREA DATASET

Model	P@1↑	P@10↑	P@50m↑	P@100m↑	mAP↑
RK-Net [48]	15.74	45.60	12.32	21.60	25.52
VIGOR [11]	38.82	**74.85**	28.95	54.50	51.31
ReIDSB [45]	29.31	57.33	21.53	36.96	38.92
X-modality [46]	34.04	55.88	36.96	41.96	41.82
DCMHN [4]	23.56	50.27	19.31	27.79	31.47
CMPC (Ours)	**57.59**	69.33	**40.79**	**70.10**	**62.76**

The bold values indicate the top performing entries.

TABLE IV
METHOD COMPARISON ON THE CROSS-AREA DATASET

Model	P@1↑	P@10↑	P@50m↑	P@100m↑	mAP↑
RK-Net [48]	4.90	21.81	3.63	7.38	10.77
VIGOR [11]	18.72	49.71	13.98	24.43	28.90
ReIDSB [45]	12.07	35.98	7.05	13.05	9.24
X-modality [46]	21.49	50.81	16.26	26.28	31.20
DCMHN [4]	14.71	40.31	11.98	18.53	23.18
CMPC (Ours)	**28.90**	**57.45**	**22.82**	**34.70**	**40.39**

The bold values indicate the top performing entries.

Regarding geolocalization methods, we use RK-Net [48] and VIGOR [11]. RK-Net is a UAV-Satellite cross-view geolocalization method that aims to classify the building from the query. We modify the instance loss with triplet loss to adapt RK-Net for the cross-modal task. VIGOR is a ground-to-satellite geolocalization method that focuses on predicting the geolocation of the query. We use neighbor patches as semipositive samples defined in VIGOR and follow the original training process to train the VIGOR model.

For cross-modal methods, we use ReIDSB [45], X-modality [46], and DCMHN [4]. ReIDSB is a strong retrieval baseline for IR-RGB person reidentification, while X-modality is a cross-modal IR-RGB retrieval method focused on modal adaptation. DCMHN is a SAR-Optical retrieval method for the classification of area categories.

In the experiment of the Aligned dataset, we compare our proposed method with previous cross-modal retrieval methods by evaluating their retrieval performance on this task, as shown in Table II. Our method achieves 81.95 and 85.98 in Top-1 precision and mAP. This improved performance can be attributed to that our method directly models the distributional gap and focuses on the instance-level retrieval objective, rather than simply extracting cross-modal features from aligned pairs. By doing so, we are able to achieve better accuracy and overcome the modal discrepancy.

In the Same-Area dataset, we further focus on the localization performance of the methods to predict the query location. Our proposed scheme achieves 57.59 and 40.79 in top-1 precision and 50-m accuracy respectively, as shown in Table III. The top part of the table lists the geolocalization retrieval methods that directly learn the embedded feature distance between query and reference patches for locating the reference image in the same area. RK-Net, which focuses on learning the position shift between query and reference patches, suffers from unstable position relations due to the gap between SAR and optical imaging modalities. VIGOR, which benefits from overlapping semipositive patches, outperforms other methods in Top-10 retrieval precision, but still suffers from modality discrepancy, resulting in unsatisfactory results in Top-1 matching precision. The bottom part of the table lists cross-modal retrieval methods designed to overcome the gap between modalities. Image transformation used in DCMHN fails to address the gap between SAR and optical modalities, and similarity reranking strategies, such as those used in X-modality, do not provide additional information to improve matching accuracy, resulting in poor performance. In contrast, our proposed CMPC scheme first overcomes the modal gap in the coarse search and then refines the matching prediction by considering the location information of retrieved candidates, achieving state-of-the-art performance in this task.

In the experiment of the Cross-Area dataset, as presented in Table IV, all methods exhibit a substantial decrease in accuracy, highlighting the difficulty of this dataset. This dataset includes nonaligned pairs between the cross-modal patches, and the test data come from a new region that did not appear in the training phase, making it more challenging than the previous two datasets. Despite this challenge, the proposed scheme still outperforms other methods in this task. However, the improvement is not as significant, as the refinement module cannot learn sufficient information in the coarse retrieval phase, which contains a vast number of outliers. Nevertheless, the proposed method still demonstrates its effectiveness in handling CMPC, even under the challenging conditions of the Cross-Area dataset.

To evaluate whether our method can also work on lower resolution data, we conduct experiments on OS-Dataset [49]. The OS-Dataset comprises 2673 nonoverlapping aligned patch pairs of 512×512 pixels with 1-m spatial resolution. We downsample the images to 200×200 pixels, making them lower than 2-m spatial resolution. When using the same image size as our proposed dataset, the low-resolution images from the OS-Dataset can capture a larger region, providing a broader field of view. Therefore, the retrieval accuracy shown in Table V is higher in general. Notably, our proposed method still outperforms other methods with a top-1 precision of 96.93 and an mAP of 97.94.

2) *Comparison With Feature Correspondence Methods*

Following the coarse search, the previous steps involve the geometric verification via features correspondence [50], [51] and correlation verification [52] to rerank putative retrieval results. Therefore, we also compare our method with these methods [52], [53], [54] on the Same-Area dataset. In this experiment, the

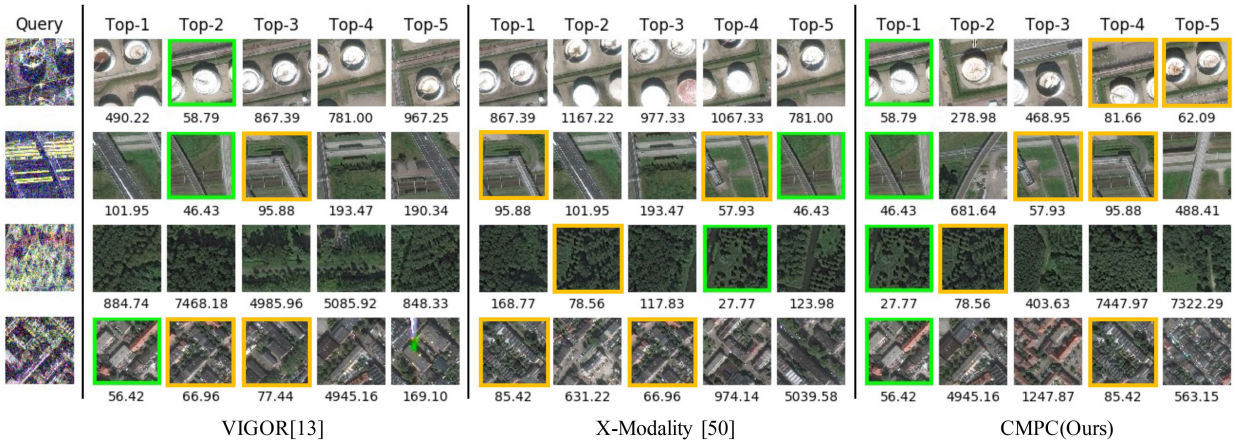

Fig. 6. Visualization of the retrieval result on the Same-Area dataset. The patches with green boxes are real matches, while the yellow boxes are neighbors with region overlaps. The numbers under patches are the location distance between the query and the retrieved optical patches.

TABLE V
METHOD COMPARISON ON THE OS-DATASET

Model	P@1↑	P@5↑	P@10↑	P@20↑	mAP↑
Dual-constrained [44]	82.08	94.10	96.46	97.88	87.49
ReIDSB [45]	81.84	93.63	95.52	97.17	86.96
DCMHN [4]	80.66	93.87	96.23	97.88	86.48
AlignGAN [47]	92.27	93.84	97.74	98.76	94.43
X-Modality [46]	92.92	95.52	96.23	96.93	94.03
CMPC (Ours)	**96.93**	**99.29**	**99.53**	**99.76**	**97.94**

The bold values indicate the top performing entries.

TABLE VI
COMPARISON OF FEATURE CORRESPONDENCE METHODS ON THE SAME-AREA DATASET

Model	P@1↑	P@5↑	mAP ↑	FPS ↑
RIFT [53]	11.90	37.90	35.46	0.16
SuperGlue [54]	16.91	58.43	33.83	3.03
CVNet [52]	31.74	59.39	44.16	7.14
CMPC (Ours)	**57.59**	**69.33**	**62.76**	**75.00**

The bold values indicate the top performing entries.

global features are extracted and the coarse search is performed. Then, we conduct a comparative analysis of these methods in the refinement stage. Notably, RIFT [53] is a multimodal image matching approach based on radiation-invariant feature descriptors, SuperGlue [54] stands out as a deep- learning-based method for feature correspondence, and CVNet [52] serves as a robust deep verification network for image retrieval tasks. In addition, we compare the computational speeds of these methods, assessed in frames per second (FPS). The results shown in Table VI demonstrate that our approach achieves superior accuracy with a higher computational speed.

3) Qualitative Analysis

We show the visualization of retrieval and match results on nonalign pairs setting and compare with the state-of-the-art instance-level retrieval methods VIGOR [11] and cross-modal methods X-modality [46]. Given the SAR queries, as shown on the left in Fig. 6, the compared methods cannot find the real match or even cannot retrieve the corresponding target in the top-ranked candidates. One reason is that they retrieve the wrong candidates due to the modal discrepancy between queries and references, resulting in matching wrong targets, which are similar in visual appearance but distinct in semantics. Another reason is that they do not improve the correspondence prediction from the retrieved candidates. Compared with previous methods, our proposed CMPC can extract more instance discriminative features; therefore, the retrieved patches have more semantic similarity and belong to the same region category. Moreover, the proposed CMPC can retrieve as most as possible neighbors in the coarse search module, which are marked as the yellow boxes, and then takes into account the location information of these retrieved candidates and adopts deep graph learning to mine the relationship between the candidates. The visualization results show that the proposed method can match the correct target from the cross-modal patch database.

TABLE VII
ABLATION STUDY OF THE MODULES

Module		Baseline	Different Configurations			
Hard mining Triplet			✓	✓	✓	✓
Feature Projector				✓	✓	✓
Wasserstein Adversarial					✓	✓
Refinement Module						✓
Same-area	Top-1	26.47	28.97	45.73	45.35	**57.59**
	mAP	32.16	34.25	55.26	54.85	**61.31**
Cross-area	Top-1	6.47	19.07	27.54	28.34	**28.90**
	mAP	8.90	27.52	37.62	38.38	**40.39**

The bold values indicate the top performing entries.

C. ABLATION STUDY

In this article, we propose the cross-modal retrieval module, the modal-invariant adversarial learning, and the refinement for the task of SAR-Optical patch correspondence. In this section, we analyze the impacts of these modules and explore their effects on different configurations to better understand their contributions to this task.

Table VII presents the performance results of our proposed methods for the Same-Area and Cross-area datasets, evaluated

TABLE VIII
COMPARISON OF DIFFERENT TYPES OF ADVERSARIAL LEARNING ON THE CROSS-AREA DATASET

Model	P@1↑	P@5↑	P@10 ↑	P@20↑	mAP ↑
Baseline [45]	56.47	84.90	90.34	95.25	70.26
Baseline w/ Classifier [1]	64.75	90.30	94.15	96.32	76.67
Baseline w/ Wasserstein (Ours)	**76.70**	**94.05**	**95.80**	**97.60**	**84.65**

The bold values indicate the top performing entries.

TABLE IX
COMPARISON OF DIFFERENT TYPES OF ADVERSARIAL LEARNING ON THE CROSS-AREA DATASET

Model	P@1↑	P@5↑	P@10 ↑	P@20↑	mAP ↑
Baseline [45]	27.54	48.33	57.45	66.43	37.62
Baseline w/ Classifier [1]	27.28	48.27	58.13	66.70	37.55
Baseline w/ Wasserstein (Ours)	**28.34**	**49.75**	**59.42**	**67.59**	**38.38**

The bold values indicate the top performing entries.

under various configurations. The baseline is the pretrained ResNet-50 supervised with the soft-margin triplet loss [34]. Compared with the soft-margin triplet, the hard-mining triplet loss can improve the performance of the retrieval module. With the proposed feature projector, the network has a gain of 16.76 points and 8.47 points on Top-1 precision on the Same-Area and Cross-Area datasets, respectively. Combined with the Wasserstein adversarial learning, the network has improved performance on the Cross-Area dataset but slightly decrease on the Same-Area dataset. By applying the refinement module, the scheme has a high performance at 57.59 points and 28.90 points on Top-1 precision on both the nonaligned datasets, respectively.

We further analyze the effectiveness of the Wasserstein learning and refinement module.

1) *Wasserstein Discriminator*

We compare our proposed Wasserstein discriminator with the traditional classifier discriminator [1] and analyze how the different types of discriminators affect retrieval performance. We conduct the experiments on the Aligned dataset and the Cross-Area dataset. The baseline is the proposed embedding network trained without adversarial learning, while the compared discriminator is a classifier discriminator, which outputs the binary modality prediction of SAR and optical.

The experiments on the Aligned dataset are shown in Table VIII, we observe that compared with the baseline, applying adversarial training with both the classifier discriminator and the Wasserstein discriminator can improve retrieval performance in the Aligned dataset. It means that the adversarial strategy can narrow the modal gap between the query and reference images when the position shift is not large. The Wasserstein discriminator has a better performance than the classifier discriminator, indicating that the Wasserstein adversarial can directly model the gap between the two modalities and achieve a narrower gap of features between SAR and optical.

We also conduct retrieval and localization performance analysis on the Cross-Area dataset, as shown in Table IX. The results indicate that the improvement on nonaligned pairs is smaller than on aligned pairs because the difference between query and reference images includes not only modal discrepancy but also positional shift. In this setting, the classifier discriminator occurs accuracy decreases. The possible explanation is that the classification-based discriminators do not learn the metric space and are not compatible with the metric-based scheme. The proposed Wasserstein adversarial learning models the modality gap in a regression way, which can have an improvement in performance than the classification-based method. Moreover, the results also imply that modal-invariant feature extraction is not enough in this task when position shifts happen, and the refinement of the coarse retrieval needs to be considered.

TABLE X
COMPARISON OF DIFFERENT TYPES OF ADVERSARIAL LEARNING ON THE OS-DATASET

Model	P@1↑	P@5↑	P@10 ↑	P@20↑	mAP ↑
Baseline [45]	81.84	93.63	95.52	97.17	86.96
Baseline w/ Classifier [1]	95.28	98.58	99.06	**99.76**	96.87
Baseline w/ Wasserstein (Ours)	**96.93**	**99.29**	**99.53**	**99.76**	**97.94**

The bold values indicate the top performing entries.

TABLE XI
COMPARISON OF CONNECTION TYPES OF THE GRAPH EDGES ON THE SAME-AREA DATASET

Connect method	K_n	K_e	P@1 ↑	mAP↑
Initial Retrieval	–	–	45.35	54.85
MLP	20	0	21.90	30.14
Full Connect	20	20	23.39	42.35
Position Embedding	20	0	55.77	62.01
Feature KNN	20	5	47.53	56.83
Location KNN(Ours)	20	5	**57.59**	**62.76**

The bold values indicate the top performing entries.

We also conduct comparison experiments on OS-Dataset [49] to evaluate the performance of different adversarial learning strategies. The experimental results shown in Table X demonstrate that our proposed Wasserstein adversarial strategy can achieve better performance than the classification adversarial one.

2) *Graph Representation for Matching Pairs*

The refinement module is a crucial part of the proposed method, which greatly improves the matching performance. We further explore the optimal configurations for different types of connections and the number of candidates and edges.

We first explore the edge connection configuration of the graph and its impact on message transmission and relationship mining, which can lead to different matching performances. Hence, we compare our graph construction method with several inlier prediction methods from the coarse retrieval result in Table XI.

1) *MLP* updates the concatenated features to predict the inlier.
2) *Full connect* connects all the nodes with the weight of the edges equal to 1, which ensures that all nodes can equally propagate messages from all other nodes.

TABLE XII
COMPARISON OF GRAPH CONFIGURATIONS ON THE SAME-AREA DATASET

K_n	K_e	P@1↑	P@50m↑	mAP↑
5	3	51.11	37.02	55.26
10	3	54.33	38.94	60.81
10	5	54.83	39.07	61.21
10	10	54.75	39.26	61.19
20	5	**57.59**	**40.79**	62.76
20	10	57.22	40.56	**62.78**

The bold values indicate the top performing entries.

3) *Position embedding* embeds the 2-D location coordinates into vectors of the same dimension as the node features using MLP and adds them to the node features.
4) *Feature KNN* connects K_e nearest neighbors with the highest inner-product between node features.

Compared with these methods, we connect K_e nearest neighbors (*Location KNN*) with the closest distance between the GPS location of candidates in a more natural way.

Compared to the coarse search, the MLP and full connect settings occur a decrease in accuracy. The results imply that the MLP cannot learn the relation between nodes, while the Full connect lost the topology discrimination. Compared to the above settings, the position embedding setting can improve performance by incorporating location information into node features. However, it does not model the spatial information into a graph structure, limiting the potential of graph networks. The Feature KNN can model feature similarity into the graph structure and improve the matching performance with the graph attention network. However, this method does not consider the location information between nodes, leading to a slight improvement in the coarse search. Our proposed Location KNN further considers the position relationship and directly models it into the graph representation, taking full advantage of the graph network. The results indicate that modeling location information directly into the graph topology can significantly improve the performance of the proposed graph refinement module.

We also conduct experiments to determine the optimal number of graph nodes K_n and the optimal number of graph edges K_e of the refinement module. The number of nodes, K_n, determines the number of candidates considered on the refinement. More nodes in the graph mean that more potential true matches are included, but it can be more challenging for the network to distinguish between these candidates for the true target. The number of edges K_e affects the message that the nodes receive from their neighbors. Increasing the number of edges leads to more knowledge being gathered into the updating nodes, while more edges can result in oversmoothing of features. Therefore, we conducted experiments to explore the optimal setting of K_n and K_e by evaluating the correspondence precision, as shown in Table XII.

The experimental results indicate that improving K_n can significantly improve the matching performance. When the candidates are not enough (K_n=5), the graph suffers from insufficient potential inliers, leading to poor matching performance. When the graph contains fewer edges (K_e=3) or is fully connected (K_n=10, K_e=10), it can neither transmit messages nor lose structural discrimination, leading to decreased accuracy. Increasing the number of candidate connections ($K_n = 20, K_e = 10$) leads to sample imbalance and makes it difficult to find positive samples as well as computation costs. In our experiments, we find out that K_n=20 and K_e=5 can make a better tradeoff in performance and computation cost.

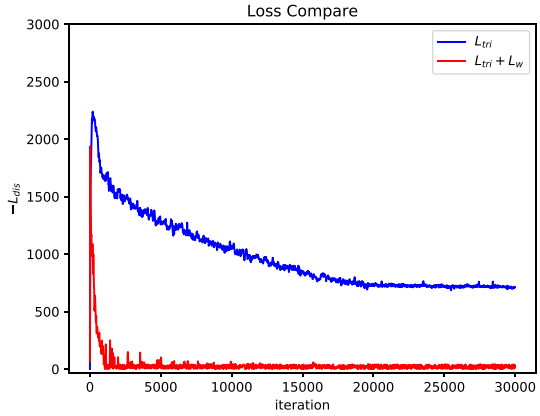

Fig. 7. Wasserstein distance is learned by the discriminator. The figure shows that utilizing only L_{tri} can partially minimize the distance but the gap still exists, and utilizing both L_{tri} and L_w can eliminate the modal discrepancy.

V. DISCUSSION

A. Wasserstein Adversarial Learning

In the proposed scheme, we design a cross-modal retrieval module based on adversarial learning with the Wasserstein discriminator. In this module, we assume that the output of the discriminator can represent the Wasserstein discrepancy between features from different modalities. During the training process, the embedding network minimizes the modal gap by both triplet loss and discrepancy loss. The triplet loss minimizes the distance of positive pairs, which are features from different modalities. As such, it can also minimize the Wasserstein discrepancy. L_w directly minimizes the discriminator's outputs, which reduces the distances between the two modalities.

To verify this assumption, we compare the Wasserstein distance, as computed by the discriminator during the training phase, under various loss functions, as illustrated in Fig. 7. The minus output of the discriminator $-L_{\text{dis}}$ represents the Wasserstein discrepancy between the feature distribution of different modalities. The triplet loss can slightly reduce this discrepancy by shortening the distance between positive samples. In contrast, the Wasserstein loss in adversarial style can directly narrow the distance. This phenomenon indicates that the Wasserstein learning strategy is effective in eliminating the gap between different modalities, highlighting the effectiveness of the proposed method.

B. Graph Representation

We present a visualization of the graph structure samples of the refinement module to demonstrate how the network learns knowledge from the graph. Fig. 8 illustrates the graph samples that can successfully estimate the true inlier. The color of the nodes represents the feature similarity between the query and

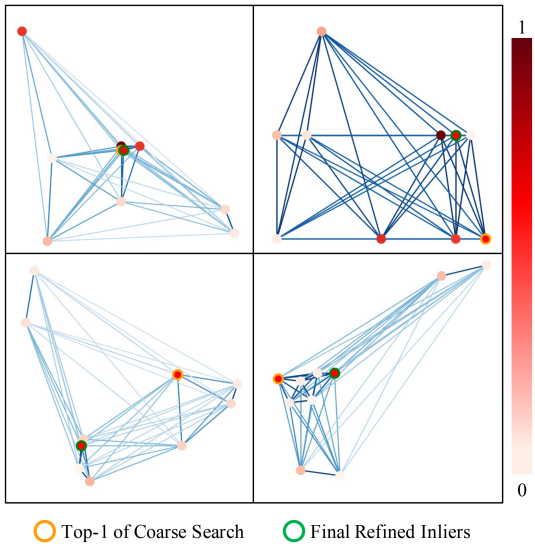

Fig. 8. Illustration of the graph samples; the green circles indicate the true positive inliers, while the orange circles indicate the top-1 retrieved item from the coarse search module. The color of the nodes indicates the feature similarity of matching pairs.

the corresponding candidate, and the color depth of the edge is calculated using (13). These graphs are then input into the attention network to predict one inlier from the nodes, whose boundary is marked in green.

Our results show that inliers tend to have closer neighbors than outliers, indicating that inliers have different graph topology distributions from the outliers. The results demonstrate that the graph neural network can mine the location relationships from the graph structure and learn to discriminate the inlier for the different distributions.

C. Limitations and Future Works

The proposed CMPC method has demonstrated promising results in the challenging scenario of SAR to optical patch correspondence. However, we have not yet explored the methods in optical to SAR correspondence scenarios due to the lack of sufficient data. Future works will focus on collecting more diverse and comprehensive datasets that cover a wider range of scenarios and modalities to comprehensively evaluate the effectiveness of the methods. Such datasets will enable us to conduct rigorous experiments and comparative analysis to establish the generalizability and robustness of the proposed method in various cross-modal correspondence scenarios.

In addition, in real-world applications, we may face situations where optical images and SAR images have very different resolutions. However, in this work, our primary focus is on addressing the cross-modal discrepancy challenge in visual localization, and we have not yet specifically provided a solution for resolution differences. We make the assumption that during the visual localization process, the spatial resolution of reference images close to the query image can be easily satisfied, thus mitigating the impact of resolution differences. We will investigate the significant resolution differences in our future work.

Finally, while the proposed method achieved promising results on the Aligned dataset and the Same-Area dataset, the performance on the Cross-Area dataset is still far from meeting real-world demands. Therefore, enhancing the robustness and generalizability of our proposed scheme is imperative for it to perform well in the wild. This could be achieved by developing techniques to handle positional shifts on the feature embedding to account for nonaligned patch pairs. Furthermore, the design of semisupervised learning for cross-modal methods could be explored to tackle the challenge of insufficient data.

VI. Conclusion

This article presented a novel coarse-to-fine scheme to tackle the challenging problem of CMPC between optical and SAR images. The proposed scheme consisted of two modules: The first of which retrieved candidate patches based on modal-invariant features, while the second module refined the retrieved candidates to identify the optimal correspondence. To enhance the cross-modal retrieval performance, we introduced the Wasserstein adversarial learning to directly model the gap between the modal distributions and train the embedding network to learn the modal-invariant features. In addition, we designed a graph representation based on the reference GPS coordinates topology to model the matching of features from coarse search and propose graph attention layers to predict the optimal correspondence from the graph representation. Through extensive experiments on three SAR-Optical patch correspondence datasets of varying difficulty levels, we demonstrated the effectiveness and superiority of our proposed method.

Acknowledgment

The authors would like to extend their sincere gratitude to the anonymous reviewers for their insightful comments and contributions to improving the quality of this article. The authors would also like to thank our colleagues for their valuable suggestions. Furthermore, the numerical calculations presented in this article were performed on the supercomputing system at the Supercomputing Center of Wuhan University.

References

[1] F. Xu, W. Yang, T. Jiang, S. Lin, H. Luo, and G.-S. Xia, "Mental retrieval of remote sensing images via adversarial sketch-image feature learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 7801–7814, Nov. 2020.

[2] S. Lowry et al., "Visual place recognition: A survey," *IEEE Trans. Robot.*, vol. 32, no. 1, pp. 1–19, Feb. 2016.

[3] H. Goforth and S. Lucey, "GPS-denied UAV localization using pre-existing satellite imagery," in *Proc. Int. Conf. Robot. Autom.*, 2019, pp. 2974–2980.

[4] W. Xiong, Z. Xiong, Y. Zhang, Y. Cui, and X. Gu, "A deep cross-modality hashing network for SAR and optical remote sensing images retrieval," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5284–5296, 2020.

[5] X. Tang et al., "Meta-hashing for remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art no. 5615419.

[6] K. Walter, M. J. Gibson, and A. Sowmya, "Self-supervised remote sensing image retrieval," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2020, pp. 1683–1686.

[7] S. Hu, M. Feng, R. M. Nguyen, and G. H. 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.

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

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

[10] 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*, 2020, pp. 1395–1403.

[11] 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.*, 2021, pp. 3640–3649.

[12] L. Jing, E. Vahdani, J. Tan, and Y. Tian, "Cross-modal center loss for 3D cross-modal retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3142–3151.

[13] F. Liu et al., "Infrared and visible cross-modal image retrieval through shared features," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 11, pp. 4485–4496, Nov. 2021.

[14] Y. Li, Y. Zhang, X. Huang, and J. Ma, "Learning source-invariant deep hashing convolutional neural networks for cross-source remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6521–6536, Nov. 2018.

[15] W. Xiong, Z. Xiong, Y. Cui, and Y. Lv, "A discriminative distillation network for cross-source remote sensing image retrieval," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1234–1247, 2020.

[16] W. Xiong, Y. Lv, X. Zhang, and Y. Cui, "Learning to translate for cross-source remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4860–4874, Jul. 2020.

[17] W. Xiong, Y. Lv, Y. Cui, X. Zhang, and X. Gu, "A discriminative feature learning approach for remote sensing image retrieval," *Remote Sens.*, vol. 11, no. 3, pp. 281–281, 2019.

[18] S. R. Dubey, "A decade survey of content based image retrieval using deep learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 2687–2704, May 2022.

[19] Y. Gong, L. Wang, R. Guo, and S. Lazebnik, "Multi-scale orderless pooling of deep convolutional activation features," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 392–407.

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

[21] C. H. Song, H. J. Han, and Y. Avrithis, "All the attention you need: Global-local, spatial-channel attention for image retrieval," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2022, pp. 2754–2763.

[22] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 499–515.

[23] 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.

[24] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *CoRR*, 2017, *arXiv:1703.07737*.

[25] Y. Sun et al., "Circle loss: A unified perspective of pair similarity optimization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6398–6407.

[26] J. Fan, Y. Wu, M. Li, W. Liang, and Y. Cao, "SAR and optical image registration using nonlinear diffusion and phase congruency structural descriptor," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5368–5379, Sep. 2018.

[27] X. Xiong, Q. Xu, G. Jin, H. Zhang, and X. Gao, "Rank-based local self-similarity descriptor for optical-to-SAR image matching," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 10, pp. 1742–1746, Oct. 2020.

[28] Q. Yu, D. Ni, Y. Jiang, Y. Yan, J. An, and T. Sun, "Universal SAR and optical image registration via a novel sift framework based on nonlinear diffusion and a polar spatial-frequency descriptor," *ISPRS J. Photogrammetry Remote Sens.*, vol. 171, pp. 1–17, 2021.

[29] A. Babenko and V. Lempitsky, "Aggregating local deep features for image retrieval," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1269–1277.

[30] S. Cui, A. Ma, Y. Wan, Y. Zhong, B. Luo, and M. Xu, "Cross-modality image matching network with modality-invariant feature representation for airborne-ground thermal infrared and visible datasets," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5606414.

[31] M. Khokhlova, V. Gouet-Brunet, N. Abadie, and L. Chen, "Cross-year multi-modal image retrieval using Siamese networks," in *Proc. IEEE Int. Conf. Image Process.*, 2020, pp. 1–5.

[32] M. Ye, J. Shen, and L. Shao, "Visible-infrared person re-identification via homogeneous augmented tri-modal learning," *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 728–739, 2021.

[33] N. Huang, J. Liu, Y. Luo, Q. Zhang, and J. Han, "Exploring modality-shared appearance features and modality-invariant relation features for cross-modality person re-identification," *Pattern Recognit.*, vol. 135, pp. 109–145, 2023.

[34] 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.*, 2019, pp. 8391–8400.

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

[36] F. Yang, Z. Wang, J. Xiao, and S. Satoh, "Mining on heterogeneous manifolds for zero-shot cross-modal image retrieval," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 12589–12596.

[37] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[38] F. Tan, J. Yuan, and V. Ordonez, "Instance-level image retrieval using reranking transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 12105–12115.

[39] L. H. Hughes, M. Schmitt, L. Mou, Y. Wang, and X. X. Zhu, "Identifying corresponding patches in SAR and optical images with a pseudo-Siamese CNN," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 784–788, May 2018.

[40] S. T. Rachev, *Duality Theorems for Kantorovich-Rubinstein and Wasserstein Functionals*. Warsaw, Poland: Instytut Matematyczny Polskiej Akademi Nauk, 1990.

[41] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 214–223.

[42] J. Shermeyer et al., "SpaceNet 6: Multi-sensor all weather mapping dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 196–197.

[43] 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/CVF Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[44] M. Ye, Z. Wang, X. Lan, and P. C. Yuen, "Visible thermal person re-identification via dual-constrained top-ranking," in *Proc. Int. Joint Conferences Artif. Intell.*, 2018, pp. 1092–1099.

[45] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 1487–1495.

[46] D. Li, X. Wei, X. Hong, and Y. Gong, "Infrared-visible cross-modal person re-identification with an X modality," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 4, 2020, pp. 4610–4617.

[47] G. Wang, T. Zhang, J. Cheng, S. Liu, Y. Yang, and Z. Hou, "RGB-infrared cross-modality person re-identification via joint pixel and feature alignment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3623–3632.

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

[49] Y. Xiang, R. Tao, F. Wang, and H. You, "Automatic registration of optical and SAR images via improved phase congruency," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 931–934.

[50] F. Radenović, A. Iscen, G. Tolias, Y. Avrithis, and O. Chum, "Revisiting Oxford and Paris: Large-scale image retrieval benchmarking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5706–5715.

[51] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[52] S. Lee, H. Seong, S. Lee, and E. Kim, "Correlation verification for image retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5374–5384.

[53] J. Li, Q. Hu, and M. Ai, "RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform," *IEEE Trans. Image Process.*, vol. 29, pp. 3296–3310, 2020.

[54] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperGlue: Learning feature matching with graph neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4938–4947.

Haoyuan Li received the B.S. degree in power engineering in 2020 from Wuhan University, Wuhan, China, where he is currently working toward the Ph.D. degree in communication and information system.

His research interests include remote sensing image processing and cross-modal visual localization.

Fang Xu received the B.S. degree in electronic and information engineering in 2018 from Wuhan University, Wuhan, China, where she is currently working toward the Ph.D. degree in communication and information system.

Her research interests include remote sensing image processing and multimodal data fusion.

Wen Yang (Senior Member, IEEE) received the B.S. degree in electronic apparatus and surveying technology, the M.S. degree in computer application technology, and the Ph.D. degree in communication and information system from Wuhan University, Wuhan, China, in 1998, 2001, and 2004, respectively.

From 2008 to 2009, he was a Visiting Scholar with the Apprentissage et Interfaces Team, Laboratoire Jean Kuntzmann, Grenoble, France. From 2010 to 2013, he was a Postdoctoral Researcher with the State Key Laboratory of Information Engineering, Surveying, Mapping and Remote Sensing, Wuhan University, where he is currently a Full Professor with the School of Electronic Information. His research interests include object detection and recognition, multisensor information fusion, and remote sensing image analysis.

Dr. Yang was the recipient of the U.V. Helava Award for the best paper in *ISPRS Journal of Photogrammetry and Remote Sensing* in 2021.

Huai Yu (Member, IEEE) received the B.S. and Ph.D. degrees in communication and information system from Wuhan University, Wuhan, China, in 2015 and 2020, respectively.

From 2018 to 2020, he has been a Visiting Ph.D. Student with Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA, where he was also a Postdoctoral Fellow from 2020 to 2021. He is currently a Research Associate Professor with the School of Electronic Information, Wuhan University. His research interests include multimodal visual feature detection and matching, structure from motion, and simultaneous localization and mapping.

Yuming Xiang (Member, IEEE) received the B.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 2013, and the Ph.D. degree in signal and information processing from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2018.

Since 2023, he has been an Associate Professor with Aerospace Information Research Institute, Chinese Academy of Sciences. His current research interests include remote sensing image registration, change detection, and 3-D reconstruction.

Dr. Xiang is the recipient of the Excellent Doctoral Thesis Award of the Chinese Academy of Sciences.

Haijian Zhang (Senior Member, IEEE) received the B.Eng. degree in electronic information engineering from Wuhan University, Wuhan, China, in 2006, and the joint Ph.D. degree in signal and information processing from the Conservatoire National des Arts et Metiers, Paris, France, and Wuhan University, in 2011.

From 2011 to 2014, he was a Research Fellow with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. He is currently an Associate Professor with the School of Electronic Information, Wuhan University. His main research interests include time–frequency analysis, array signal processing, and multimedia forensics and security.

Gui-Song Xia (Senior Member, IEEE) received the Ph.D. degree in image processing and computer vision from CNRS Le Laboratoire de Traitement et Communication de l'Information, Télécom ParisTech, Paris, France, in 2011.

From 2011 to 2012, he was a Postdoctoral Researcher with the Centre de Recherche en Mathématiques de la Decision, CNRS, Paris-Dauphine University, Paris, for one and a half years. He is currently a Full Professor in Computer Vision and Photogrammetry with Wuhan University, Wuhan, China. He was also a Visiting Scholar at DMA, École Normale Supérieure, Paris, for two months in 2018. He is also a Guest Professor with Future Lab AI4EO, Technical University of Munich, Munich, Germany. His current research interests include mathematical modeling of images and videos, structure from motion, perceptual grouping, and remote sensing image understanding.

Dr. Xia is on the Editorial Board of several journals, including *ISPRS Journal of Photogrammetry and Remote Sensing*, *Pattern Recognition, Signal Processing: Image Communications*, *EURASIP Journal on Image and Video Processing*, *Journal of Remote Sensing*, and *Frontiers in Computer Science: Computer Vision*.