A Discriminative Distillation Network for Cross-Source Remote Sensing Image Retrieval

Wei Xiong, Zhenyu Xiong, Yaqi Cui, and Yafei Lv

***Abstract*—Nowadays, several remote sensing image capturing technologies are used ranging from unmanned aerial vehicles to satellites. Powerful learning-based discriminative features play an essential role in content-based remote sensing image retrieval (CBRSIR). Cross-source CBRSIR (CS-CBRSIR) is used to find relevant remote sensing images across different remote sensing sources (i.e., multispectral images and panchromatic images). But it is limited by large cross-source and intrasource variations caused by different semantic objects, spatial resolution, and spectral resolution. The main limitation of CS-CBRSIR is that it cannot address the inconsistency between different sources and exploit the intrinsic relation between them. This study proposes a discriminative distillation network for CS-CBRSIR to address this limitation. To enlarge the interclass variations and reduce the intraclass differences, the discriminative features from the first source are first extracted with a well-designed joint optimization configuration (JOC) on the basis of deep neural networks. Thereafter, the features extracted from the first source are used as a supervision signal for the second source; feature distribution in common feature space between the first and second sources are made significantly similar. The method proposed in this study simultaneously handles the cross-source and intersource variations, unlike the existing methods. Extensive experiments on the DSRSID dataset with Euclidean distance verify the effectiveness of our proposed method.**

***Index Terms*—Cross-source content-based remote sensing image retrieval (CS-CBRSIR), discriminative features, distillation network, joint optimization configuration (JOC).**

I. Introduction

THE rapid development in aerial vehicle technologies and remote sensing sensors has led to the rapid growth of remote sensing images both in quantity and quality. We have entered an era of remote sensing big data (RSBD) [1], [2]. Currently, effective management, mining, and interpretation of remote sensing images constitute crucial current research areas that need urgent solutions [53]–[56]. Consequently, useful information must be automatically drawn from RSBD. To make full use of RSBD, content-based remote sensing image retrieval (CBRSIR) [3], [4] has increasingly attracted research interests following its broad applications on retrieving interesting images according to their visual content.

Fig. 1. Intraclass variations. Images of the same class (high building) captured by panchromatic and multispectral sources [5].

Most of the recent methods focus only on unified-source CBRSIR (US-CBRSIR) [20]–[23], where the inquiry image and retrieved images are from the same data source. The development of remote sensing technologies has significantly enhanced the ability of remote sensing image acquisition. Satellites possess the ability to use two sensors to capture multispectral and panchromatic images separately. This article focuses on deep neural networks for cross-source CBRSIR (CS-CBRSIR) in which the inquiry and retrieved images are from different data sources. Notably, there has been no effective method for the CS-CBRSIR issue until the proposal of source-invariant deep hashing convolutional neural networks (SIDHCNNs) [5], which allow sensing between panchromatic and multispectral sources. Panchromatic images possess low spectral resolution but high spatial resolution, while the opposite characterizes multispectral images. Finding a feature extractor for obtaining discriminative features from different data sources poses a major issue in CS-CBRSIR. However, the current methods limit the retrieval performance due to drastically increasing volume and complexity of RSBD. Significant intraclass variations make CBRSIR [5] more complicated (Fig. 1).

Numerous methods, such as distance metric learning [16], attention mechanism [22], and multitask learning [22], have been introduced to learn a discriminative metric space and feature representation in an attempt to solve these limitations. However, these methods only address the US-CBRSIR problems, and cannot be effectively extended to CS-CBRSIR as there is a considerable data drift across sources. This is because data from panchromatic and multispectral images are heterogeneous. Specifically, CS-CBRSIR suffers from significant cross-source hard negatives $(D(b, A) < D(a, A))$ and large intrasource variations $(D(C, D) < D(C, C'))$ (Fig. 2). The existing method

Manuscript received December 15, 2019; revised February 7, 2020; accepted March 11, 2020. Date of publication March 23, 2020; date of current version April 13, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61790550, Grant 61790554, and Grant 91538201. *(Corresponding authors: Zhenyu Xiong.)*

The authors are with the Research Institute of information Fusion, Naval Aviation University, Yantai 264001, China (e-mail: xiongwei@csif.org.cn; x_zhen_yu@163.com; cui_yaqi@126.com; yfei_lv@163.com).

This article has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

Digital Object Identifier 10.1109/JSTARS.2020.2980870

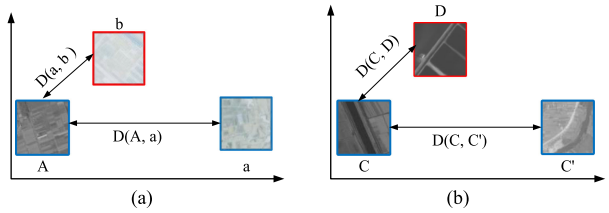


Fig. 2. Large intraclass variations and small interclass dissimilarity caused by (a) Cross-source and (b) Intrasource variation. The box color represents the class of the image [5].

cannot simultaneously address cross-source and intrasource variations.

To address these issues, this study proposes a discriminative distillation network method for CS-CBRSIR, which exploits the intrinsic relationship between panchromatic and multispectral images in a sequential manner. The method adopts the knowledge distillation capability of transferring supervision signal from one data source to the next. The first step involves the extraction of discriminative features using a new joint optimization configuration (JOC) from the first source. JOC combines three losses with a batch normalization (BN) layer to address the issue of intraclass inconsistency and interclass indistinction. Second, using the guidance of the features and weights from the first source, features for the second source are extracted. In the training process, weights of the first network F initialize the second network G facilitating the transfer of the supervision signal to network G. Additionally, all the midlevel to high-level layers in the training of the network G are frozen, while retaining the first source's high-level semantic features and learning important low-level features from the second data source. The features extracted from the first source images through the frozen network F are taken as the groundtruth for the images from the second source. Besides, the parameters of low-level layers are independent in extracting source-specific features in this two steps distillation network, which successfully addresses the cross-source issue caused by data drift. This proposed method aims to make samples with the same label very closely to each other and samples with different labels separately far apart in the common feature space for both sources (Fig. 3).

The main contributions of this article can be summarized as follows.

1) To the best of our knowledge, this is the second study exploring CS-CBRSIR after [2], which is proposing the possibility and potential values of CS-CBRSIR. Moreover, the study offers a new framework for transforming cross-source images into unified-source images for CS-CBRSIR.
2) The study proposes a cross-source distillation network to address the problem of CS-CBRSIR. This distillation network successfully solves the challenge of data drift. Additionally, this is a general model method for CS-CBRSIR that can be extended to other sources and modalities.
3) To enhance the discriminative power, we design a novel JOC. By combining JOC and the distillation network, the intrasource and cross-source variations are simultaneously solved.

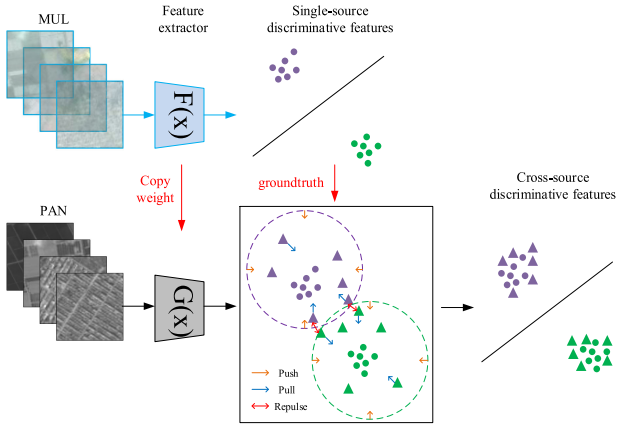


Fig. 3. Intuition of the proposed method. The ball represents the feature of multispectral image and the triangle represents the feature of panchromatic image. The color represents the class of the image [5].

The rest of this article is organized as follows. Section II presents some published work related to US-CBRSIR, cross-model retrieval in remote sensing and knowledge distillation. Our proposed method is presented in Sections III, and IV displays the experimental results and analysis in detail. Section V draws some conclusions.

II. Related Work

A. Unified-Source Content-Based Remote Sensing Image Retrieval (US-CBRSIR)

Designing an effective feature extractor for the US-CBRSIR stands out as its key challenge. Existing feature extractors can be divided into three categories based on the feature type: low-level, midlevel, and high-level feature extractors. The low-level feature extractors rely on handcrafted features, whose design requires sufficient domain expertise and engineering skills. This feature is widely exploited in CBRSIR works. Popular handcrafted features include global features constituting spectral (colour) [6], [34], texture [7], [35], shape [8], [36] as well as local features based on scale invariant feature transform [9], difference of Gaussian [38], and speeded up robust features [10]. These low-level features do not efficiently represent highly complex remote sensing images, which limit the retrieval performance. On the contrary, midlevel feature extractors produce more discriminative features by aggregating local features, such as bag-of-words [11], vector of locally aggregated descriptors [12], and Fisher vector [13] or their variants. Similarly, the features produced by midlevel feature extractors constitute handcrafted features. Considering that the same class of remote sensing images might have different orientations, scales and illuminations, the midlevel level features are not effective in describing the rich information of the images accurately. The "semantic gap" between low-level features and high-level semantic information make the handcrafted features difficult to express image semantics precisely [39]. Numerous works [14], [15] have attempted to use the convolutional neural network (CNN) in extracting highly abstractive and high-level semantic information features, which have been shown to possess superior performances to

traditional handcrafted features in CBRSIR in addressing this difficulty. In [40], the authors conduct the usual training from scratch. However, remote sensing dataset cannot provide enough data for CNNs training from scratch. Nevertheless, with the development of transfer learning, extracting feature from pre-trained networks directly [17], [18], [52] and fine-tuning of pre-trained model [19], [47], [51] on massive datasets have been proven to be more efficient and effective. The targets of remote sensing images are smaller and more complex compared with natural images. Of note, this model trained on natural images cannot be directly used in remote sensing images due to various adaptability and transferability factors. Numerous studies propose the acquisition of more discriminative features, [20] aggregate deep local features [21], and a novel region-wise deep feature representation for US-CBRSIR. For instance, [22] study shows the superiority of attention mechanism and multitask learning on US-CBRSIR tasks. In [23], the author proposed a deep hashing neural network of solving the issue of large-scale US-CBRSIR. The semisupervised graph-theoretic method [24] and region convolution features [37] are used for multilabel US-CBRSIR. While these methods can address the US-CBRSIR problems adequately, they cannot be effectively extended to CS-CBRSIR.

B. Cross-Model Retrieval in Remote Sensing

It is well known that cross-model retrieval has received widespread attention in recent years. Specially, CS-CBRSIR is an important branch of cross-modal retrieval in remote sensing. However, regarding the scarcity of remote sensing databases, there are only three works about the cross-modal retrieval problem. These three studies introduce the pioneer cross-modal retrieval works, allowing the model between remote sensing images and spoken audio, remote sensing images and sentences and panchromatic and multispectral images. In [25], the authors propose a novel deep visual-audio network (DVAN) for the retrieval of spoken audio and remote sensing image on their own dataset. This method confirms the feasibility of visual-to-image retrieval, and presents a new way of image retrieval. However, the cross-modal retrieval between audio and image differs from the cross-source retrieval between different images. Remote sensing image has more semantic information than audio signal, and these two are not significantly correlated. Thus, the model cannot be directly applied to CS-CBRSIR. Additionally, a large-scale benchmark dataset RSICD including various remote sensing images and sentences are presented in [26]. To describe the remote sensing images with accurate and flexible sentences, the authors consider some special characteristics such as scale ambiguity, category ambiguity, and rotation ambiguity while annotating the sentences in remote sensing captioning. A comprehensive review of popular caption methods is presented on their dataset to evaluate the validation. But the structure of text feature extraction cannot be applied to image features, it is hard to use this model to solve the CS-CBRSIR task. Based on deep cross-modal hashing [27], a SIDHCNNs [5] which contain a series of optimization constraints are proposed for CS-CBRSIR. While this method loses the exclusive source-specific characteristics by using a similar architecture to map image features from different sources into one common space at one training stage.

C. Knowledge Distillation

Knowledge distillation involves the transfer of knowledge from a cumbersome teacher network to a lightweight student network. This area has been widely studied in recent years. For instance, [28] proposes that a distillation network compresses vital information from an already trained teacher network into a student network. Concerning the RGB and depth cross-model task, a distillation network constitutes the transfers of knowledge between RGB and depth images of the same scene [29]. Recent studies explore different forms of "knowledge." For example, [30] indicates that knowledge distillation involves the transfer of teacher network to student network using gradient-based and activation-based spatial attention mechanism. Experimental results demonstrate that the approach significantly improves performance across several datasets. Unlike the previous knowledge distillation methods, which only regarded the output layer of a teacher network as the goal of the student's mimic, [31] focuses on the frame-per-second matrix generated by the inner product between features from two layers. Given sufficient initial weight, the parameter can rapidly reach the global optimum and improve the performance of a small work. In metric learning, the transferred knowledge is the cross-sample similarities, and the "learn to rank" technique is used in metric learning formulation [32]. This method can be widely used in numerous fields, for instance, image retrieval, face recognition, person re-identification (Re-id), and image clustering. In recent years, generative adversarial network has made impressive progress. For example, the application of learning loss approach through conditional adversarial networks to transfer knowledge from teacher to student successfully optimizes the knowledge distillation strategy [33]. However, these methods are mainly used to solve the problems of Re-id and data imbalance in natural images. Currently, none has used knowledge distillation to solve the CS-CBRSIR problem.

Inspired by these studies, we adopt a cross-source distillation network into the CS-CBRSIR. Contrary to previously proposed methods that mainly focused on training the model in a parallel manner [5], this study's method aims to train the networks in a sequential way to protect and explore source-specific features of panchromatic and multispectral images.

III. Proposed Method

This section describes the processes of our discriminative distillation network, including JOC and cross-source knowledge distillation. Different from existing works in [5], our model is trained in two stage, thus the training of images from different sources is performed separately. The main difference is that we train our model in a sequential manner, not in parallel. The framework of the proposed method is shown in Fig. 4. First, Section III-A introduces the discriminative features extracted from the first source with JOC. Second, the cross-source knowledge distillation is conducted in Section III-B to transfer the

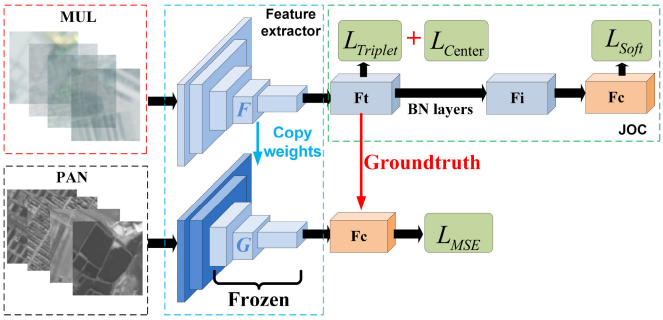


Fig. 4. Framework of our proposed cross-source distillation network.

knowledge from the first source to the second source, mapping the feature to the common space.

A. Joint Optimization Configuration (JOC)

The first step of knowledge distillation involves the training of the single-source network. It constitutes the design of a new JOC to aid in the extraction of more discriminative features from single-source images. After that, the network *F* is frozen for the second step of training.

Generally, cross-entropy is a commonly used loss function in multiclassification problems to divide the feature space into different subspaces by constructing several hyperplanes. The features of different classes will be distributed in different subspaces. The softmax loss function is presented as

$$L_{\text{Soft}} = -\sum_{i=1}^{m} \log \frac{e^{W_{y_i}^T F_c(x_i)+b_{yi}}}{\sum_{j=1}^{n} e^{W_j^T F_c(x_i)+b_j}} \tag{1}$$

where m is size of mini-batch and n is the number of class. x_i denotes the input image. $F_c(x_i) \in R^d$ denotes the network output before the softmax function, belonging to the y_ith class. $W_j \in R^d$ denotes the jth column of the weights and $b \in R^n$ is the bias term.

However, the classification task is different from the retrieval task as the deeply learned features with softmax tend to be separable rather than discriminative. Therefore, these features cannot be directly used for retrieval. We construct the triplet loss [41], to improve the discriminative power of deep features and enhance the interclass dispersion and intraclass compactness. However, the interclass distance is sometimes smaller than intraclass due to the particularity of remote sensing image. In such a semantic space, it is hard to guarantee optimal global constraints. The convergence speed on solving some problems with high complexity will be significantly slowed down. We can combine softmax loss and triplet loss to train the model, thus enabling the training process to learn more discriminative features and easier to converge. In the training process, given an x_i^a (anchor) of a special class, we ensure that an image from the same class x_i^p (positive) is closer to the anchor than that of an image x_i^n(negative) belonging to any other class. The triplet loss is defined as

$$L_{\text{Triplet}} = \sum_{i=1}^{m} \left[d\left(F_t\left(x_i^a\right), F_t\left(x_i^p\right)\right) - d\left(F_t\left(x_i^a\right), F_t\left(x_i^n\right)\right) + \alpha \right] \tag{2}$$

where $F_t(x_i) \in R^d$ denotes the network output before the BN layer. d denotes the similarity metric. α is a margin that is enforced between positive and negative pairs.

Although the combination of triplet loss and softmax loss can improve the discriminant of the deep learned feature, some shortcomings cannot be neglected. Triplet loss considers only the difference between $d(F_t(x_i^a), F_t(x_i^p))$ and $d(F_t(x_i^a), F_t(x_i^n))$, while totally ignoring the absolute values. The triplet loss is determined by two classes of images sampled randomly. It is hard to ensure that $d(F_t(x_i^a), F_t(x_i^p)) < d(F_t(x_i^a), F_t(x_i^n))$ is maintained throughout the training process.

Center loss [42], aims at learning a center of each class and minimize the distances between the features and their corresponding class centers simultaneously. The combination of center loss and triplet loss make up for the drawbacks of only employing triplet loss. The center loss function is given in (3)

$$L_{\text{Center}} = \frac{1}{2} \sum_{i=1}^{m} \left\| F_t\left(x_i\right) - c_{y_i} \right\|_2^2 \tag{3}$$

where $c_{y_i} \in R^d$ denotes the y_i th class center of deep feature. The JOC totally includes three losses as follow:

$$L_{\text{Total}} = L_{\text{Soft}} + L_{\text{Triplet}} + \lambda L_{\text{Center}} \tag{4}$$

where λ is hyperparameter, balancing weight of center loss.

Finally, with JOC, the interclass differences are enlarged, while the intraclass variations are reduced. Besides, we add a BN layer after features. The extracted feature F_t is transformed into the normalized feature F_i after passing through the BN layer. To compute softmax loss, and center loss and triplet loss in the training process, F_i and F_t are used, respectively. Normalization makes each dimension of F_i more balanced. The features are likely to have a Gaussian distribution with the aim of making the softmax loss easier to converge. Besides, the introduction of the BN layer significantly reduces the constraints of the softmax loss and makes the triplet loss easier to converge, simultaneously. This is mainly because BN force the more biased distribution back to the standard distribution, so that the input value falls in the area where the nonlinear function is more sensitive to input. In this way, small change of input leads to large changes in the loss function. Furthermore, it keeps the feature distribution of the same class compact by normalizing layer input.

B. Cross-Source Knowledge Distillation

This step involves the extraction of features on the cross-source images through the distillation network. The weight of the network *F* is transferred to the network *G* which is dedicated to the second source. Additionally, the low-level features for panchromatic and multispectral image that are most informative for this task are distinct. Hence, the weights of the network are frozen from the midlevel layer to the high-level layers. This

retains the high-level semantic features from the first source, while the second network learns the meaningful low-level features from the second source. Meanwhile, the higher layers are fixed as network still needs to map its input to a particular output representation, namely the common feature space established during training the first network. In this way, we end up with two networks, in which the lower layers extract source-specific features, and the higher layers map to the common feature space to enable cross-source retrieval.

We utilize images x_i^{s1} from the first source and x_i^{s2} from the second source in training the distillation network. The purpose of this training is to make the first source's images x_i^{s1} with label y as close to the second source's images x_i^{s2} with label y in the feature space. The features extracted from the first source images through the frozen network F are taken as the groundtruth for the second source's images. At the same time, the distillation network G is trained using the second source's images. Using the mean squared error (MSE) loss between the features of paired images $F(x_i^{s1})$ and $G(x_i^{s2})$, the feature of two sources is mapped into a common space. MSE loss is defined as

$$L_{\text{MSE}} = \sum_{i=1}^{m} \left\| G\left(x_i^{s2}\right) - F\left(x_i^{s1}\right) \right\|^2. \tag{5}$$

IV. Experiments and Analysis

To verify the effectiveness of our method, an extensive series of experiments has been considered to validate the proposed discriminative distillation network. Section IV-A introduces the experimental setup and evaluation criteria; the effectiveness of proposed JOC is verified in Section IV-B; Section IV-C presents the validity of knowledge distillation with JOC; Section IV-D analyzes the impacts of parameter λ and frozen layers on the results with the cross-source retrieval; and Section IV-E presents the comparison results with some baselines.

A. Experimental Setup

Experiments are performed on the only publicly available CS-CBRSIR datasets DSRSID [5]. The dataset comprises a great quantity of pairs of panchromatic and multispectral images which are acquired by GF-1 multispectral sensors and GF-1 panchromatic sensors, respectively. DSRSID includes 80 000 pairs of multisource images of eight classes, including aquafarm, cloud, forest, high building, low building, farm land, river, and water. Each class of image contains 10 000 pairs of multisource images. The image size of multispectral image is 64 × 64 with a resolution of 8 m, and the number of spectral channels is 4. While the image size of panchromatic image is 256 × 256 with a spatial resolution of 2 m, the number of spectral channels is only 1. Some pairs of examples from the DSRSID [5] are visually shown in Fig. 5. In this study's experiments, DSRSID is split into two subsets for constructing the training set $D_{\text{Train}} = \{(P_i, M_i, L_i)|i = 1, 2, \ldots, N\}$ and the testing set $D_{\text{Test}} = \{(P_i, M_i, L_i)|i = 1, 2, \ldots, Q\}$, where P_i denotes the panchromatic image, M_i denotes the multispectral image, and L is the image's label. N is the volume of the training data set, and Q is the volume of the testing data set. N and Q are set to 75 000 and 5000, respectively.

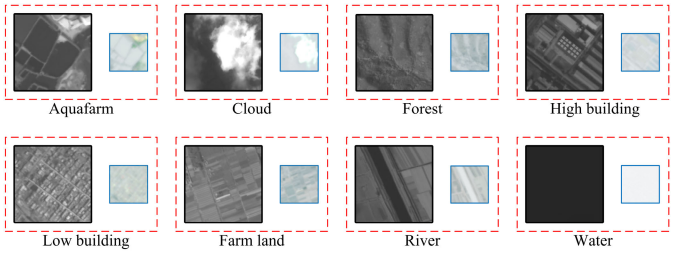


Fig. 5. Examples from DSRSID [5].

TABLE I
Architecture of Distillation Network

Layer name	Output size	18-layer	50-layer
Conv1	128*128	7*7, 64, stride2	
Conv2	64*64	3*3 maxpool, stride2	
		$\begin{bmatrix} 3*3, 64 \\ 3*3, 64 \end{bmatrix}$ *2	$\begin{bmatrix} 1*1, 64 \\ 3*3, 64 \\ 1*1, 256 \end{bmatrix}$ *3
Conv3	32*32	$\begin{bmatrix} 3*3, 128 \\ 3*3, 128 \end{bmatrix}$ *2	$\begin{bmatrix} 1*1, 128 \\ 3*3, 128 \\ 1*1, 512 \end{bmatrix}$ *4
Conv4	16*16	$\begin{bmatrix} 3*3, 256 \\ 3*3, 256 \end{bmatrix}$ *2	$\begin{bmatrix} 1*1, 256 \\ 3*3, 256 \\ 1*1, 1024 \end{bmatrix}$ *6
Conv5	8*8	$\begin{bmatrix} 3*3, 512 \\ 3*3, 512 \end{bmatrix}$ *2	$\begin{bmatrix} 1*1, 512 \\ 3*3, 512 \\ 1*1, 2048 \end{bmatrix}$ *3
	1*1	Global average pool	
Ft	1*1	512	2048
Fi	1*1	512	2048

In the experiment, the Euclidean distance is used for the similarity measure during the retrieval stage. The channel of a panchromatic image is copied into four input channels to create a 4-channel image, similar to multispectral. Each panchromatic image is resized into 256 × 256 pixels. The shallower network, Resnet18 [44], and a deeper network, Resnet50 [4] are chosen as the feature extractors. The architecture of our distillation network is provided in Table I. The Adam optimizer with a learning rate of 0.001 and batch size of 128 and 32 for Resnet18 and Resnet50, are considered, respectively. The margin α of triplet loss is set to 0.3. For quantitative evaluation, the precision at k (k is the number of returned images) is reported along with the mean average precision (mAP).

TABLE II
QUANTIFYING THE EFFECTIVE OF DIFFERENT OPTIMIZATION CONFIGURATIONS ON MUL RETRIEVAL TASK

Feature Extractor	Loss	P@1	P@3000	P@8000	mAP
Resnet18	Softmax	0.9157	0.8503	0.7820	0.8142
	Triplet	0.9412	0.9108	0.8839	0.9012
	Triplet+Softmax	0.9715	0.9417	0.9128	0.9613
	Triplet+Center+Softmax	0.9920	0.9813	0.9872	0.9898
Resnet50	Softmax	0.9215	0.9367	0.8100	0.8421
	Triplet	0.9572	0.9498	0.9023	0.9314
	Triplet+Softmax	0.9813	0.9802	0.9698	0.9713
	Triplet+Center+Softmax	0.9954	0.9901	0.9958	0.9917

TABLE III
QUANTIFYING THE EFFECTIVE OF DIFFERENT OPTIMIZATION CONFIGURATIONS ON PAN RETRIEVAL TASK

Feature Extractor	Loss	P@1	P@3000	P@8000	mAP
Resnet18	Softmax	0.8836	0.8325	0.7328	0.7979
	Triplet	0.9021	0.8612	0.8319	0.8529
	Triplet+Softmax	0.9417	0.9208	0.9036	0.9206
	Triplet+Center+Softmax	0.9813	0.9725	0.9802	0.9762
Resnet50	Softmax	0.9013	0.9215	0.7831	0.8210
	Triplet	0.9123	0.9218	0.8563	0.8792
	Triplet+Softmax	0.9762	0.9692	0.9431	0.9643
	Triplet+Center+Softmax	0.9802	0.9903	0.9810	0.9892

B. Effective of Joint Optimization Configuration (JOC)

For performance evaluation with individual sources (MUL and PAN separately), several neural network optimizations have been investigated. Since this section intends to intuitively verify whether our method can help to extract more discriminative features on single-source images, four major optimization configurations are considered: only the adoption of softmax loss; only the adoption of triplet loss; the combination of triplet and softmax loss; and the combination of triplet, center and softmax loss. These four optimizations are compared and the corresponding results evaluated further on two single-source retrieval tasks. Under these different optimization configurations, the proposed method achieves the best performance under the same network architecture for two single-source retrieval tasks (Tables II and III). The performance shows that training with JOC improves the discrimination of learned features for two sources. In cases of similar optimization configurations, the results improve as the network deepens. Additionally, the results imply that multispectral retrieval task is relatively superior to the panchromatic retrieval task. This is because, in addition to focusing on the panchromatic images' spatial details, the method also makes full use of the existing information, including spectral details in the multispectral images.

Furthermore, under the same network architecture, Resnet50, the feature distributions under various optimization configurations are visualized by the method of t-distributed stochastic neighbor embedding (t-SNE) [43] (Fig. 6). This intuitively exhibits the feature distributions of the multispectral and panchromatic images under softmax loss, triplet loss, triplet loss with softmax loss, the combination of softmax loss, triplet loss and center loss, respectively. The distribution of the same class in Fig. 6(d) and (h) is more compact and its features relatively separable compared with Fig. 6(a)–(g). In conclusion, the features trained under proposed optimization configuration are more discriminative since its more dispersed interclass and more compact intraclass on two single-source retrieval tasks. This verifies that training with JOC solves significant intrasource variations.

C. Validity Analysis of Knowledge Distillation With JOC

The networks for single-source were analyzed in the previous IV-B section. These networks correspond to the training of the cross-source distillation. The purpose of the distillation network designed in this study is to extract discriminative features of cross-source images, but those with similar labels should be close to each other in the feature space, while images with different labels should be distinctively placed.

Tables IV and V show the results of the distillation networks trained on the cross-source retrieval tasks between multispectral and panchromatic images. Unlike single-source retrieval, the image source of query is totally different from the image source of gallery on cross-source retrieval. Results are shown for the

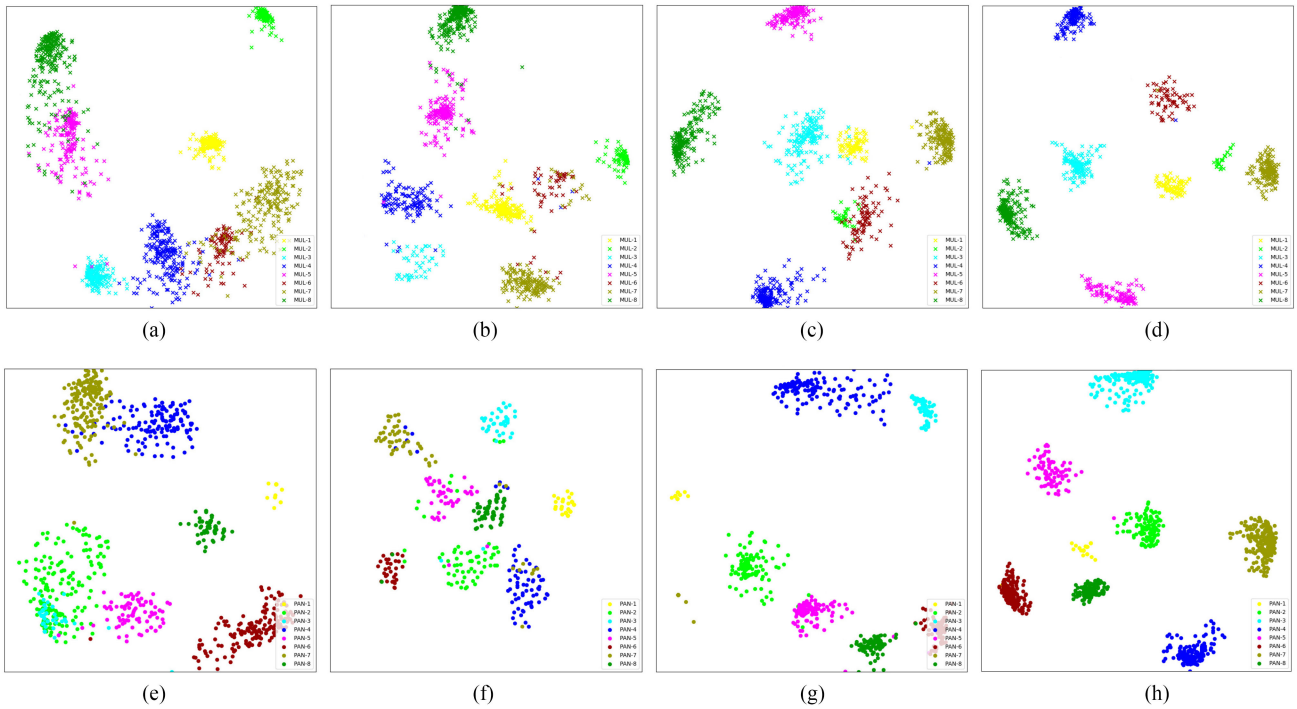


Fig. 6. Feature visualization of the learned features on different single-source retrieval tasks: (a) MUL, Softmax. (b) MUL, Triplet. (c) MUL, Triplet+Softmax. (d) MUL, Triplet+Center+Softmax. (e) PAN, Softmax. (f) PAN, Triplet. (g) PAN, Triplet+Softmax. (h) PAN, Triplet+Center+Softmax.

TABLE IV
QUANTIFYING THE EFFECTIVE OF DIFFERENT OPTIMIZATION CONFIGURATIONS ON MUL->PAN RETRIEVAL TASK

Feature Extractor	Loss	P@1	P@3000	P@8000	mAP
Resnet18	Softmax	0.9032	0.8479	0.7718	0.8069
	Triplet	0.9392	0.9012	0.8617	0.8998
	Triplet+Softmax	0.9702	0.9368	0.9019	0.9589
	Triplet+Center+Softmax	0.9897	0.9698	0.9786	0.9701
Resnet50	Softmax	0.9160	0.9294	0.8026	0.8375
	Triplet	0.9478	0.9302	0.8901	0.9085
	Triplet+Softmax	0.9787	0.9786	0.9637	0.9686
	Triplet+Center+Softmax	0.9898	0.9798	0.9714	0.9811

TABLE V
QUANTIFYING THE EFFECTIVE OF DIFFERENT OPTIMIZATION CONFIGURATIONS ON PAN -> MUL RETRIEVAL TASK

Feature Extractor	Loss	P@1	P@3000	P@8000	mAP
Resnet18	Softmax	0.8713	0.8247	0.7295	0.7902
	Triplet	0.8819	0.8601	0.8231	0.8400
	Triplet+Softmax	0.9312	0.9076	0.8893	0.9167
	Triplet+Center+Softmax	0.9769	0.9658	0.9601	0.9679
Resnet50	Softmax	0.8902	0.9146	0.7789	0.8106
	Triplet	0.9011	0.9214	0.8491	0.8682
	Triplet+Softmax	0.9689	0.9601	0.9315	0.9568
	Triplet+Center+Softmax	0.9781	0.9875	0.9702	0.9798

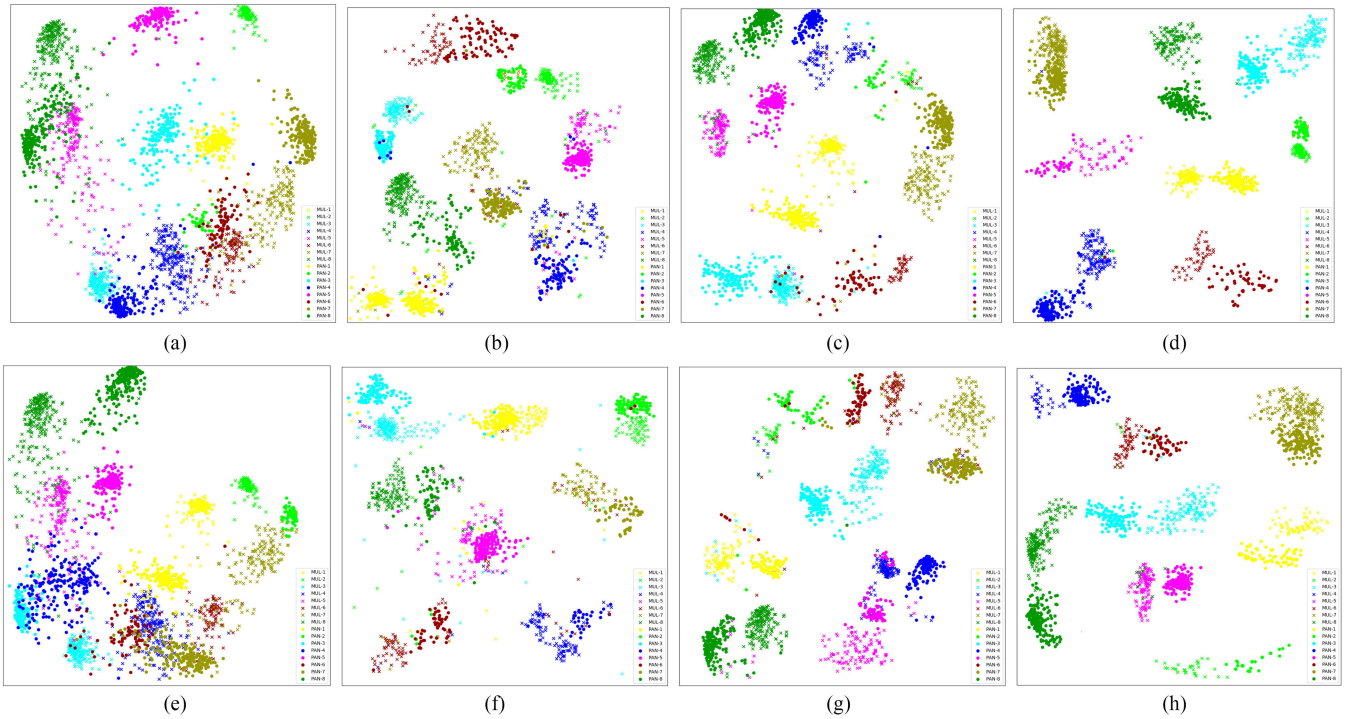


Fig. 7. Feature visualization of the learned features on different single-source retrieval tasks: (a) MUL->PAN, Softmax. (b) MUL->PAN, Triplet. (c) MUL->PAN, Triplet+Softmax. (d) MUL->PAN, Triplet+Center+Softmax. (e) PAN -> MUL, Softmax. (f) PAN -> MUL, Triplet. (g) PAN -> MUL, Triplet+Softmax. (h) PAN -> MUL, Triplet+Center+Softmax.

two feature extractor architectures Resnet18 and Resnet50 with four optimization configurations. Compared with the results on the single-source retrieval, distillation network's results decline slightly on the two cross-source retrieval tasks. Therefore, the trained distillation network significantly retains the high-level semantic features from the first source as well as learns the meaningful low-level features from the second source (Tables IV and V). Based on JOC's superiority, the proposed method could achieve the best performance under the same network architecture. The accuracy of transferring from multispectral images to panchromatic images is slightly higher than from panchromatic images to multispectral images. Furthermore, the feature distributions acquired on two cross-source datasets are visualized by the method of t-SNE [43] (Fig. 7), which intuitively shows that the problem of data drift is further solved using distillation network.

D. Parameter Analysis

In this section, experiments are conducted to illustrate the effect of parameter λ on the performance of the proposed method. Tables VI and VII show the results of two network architectures with different hyperparameters λ on multispectral and panchromatic images, respectively. In the experiments, the parameter λ is set to 0.00005, 0.0005, 0.005, 0.05, and 0.5. The optimal performance is acquired when λ is set to 0.0005 on multispectral retrieval task and λ is set to 0.005 on panchromatic retrieval task. The performance decreases sharply as λ continues to increase. With an appropriate λ, the discriminative power of deep features is significantly enhanced. Therefore, the proposed method significantly improves discriminative power of deeply learned features.

As mentioned above, λ is an important parameter involved with performance of single-source retrieval task. Thus, we focus on the analysis the effect of parameter on the cross-source retrieval tasks. More specifically, Tables VIII and IX report the results of different network architectures with different λ on cross-source MUL$\rightarrow$PAN retrieval task and cross-source PAN$\rightarrow$MUL retrieval task, respectively. It is clear that the best performance is acquired when λ is set to 0.0005 on MUL$\rightarrow$PAN retrieval task and λ is set to 0.005 on PAN$\rightarrow$MUL retrieval task. Similar to single-source retrieval tasks, the discriminative power of deep features can be significantly enhanced with proper λ.

The parameter experiments are conducted on different margins α in (2) for different retrieval tasks (Fig. 8). It can be seen that the best mAP value is achieved when the α is set to 0.3, and the results becomes lower with the α increases. This is mainly because the large margin makes the triplet loss maintain a large value, making it less likely to approach 0. And a small margin makes the triplet loss easy to approach 0; the trained feature does not have a good discrimination ability. Therefore, it is crucial to set a reasonable margin value.

Besides, the cross-source distillation method is highly dependent on the successful knowledge transfer from MUL$\rightarrow$PAN or PAN$\rightarrow$MUL. We evaluated the influence on network accuracy in the cross-source tasks with varying components for knowledge transfer to gain more insights into this transfer. Tables X and XI, and Fig. 9 show the impact of freezing of different

TABLE VI
QUANTIFYING THE EFFECTIVE OF DIFFERENT LAMBDA ON MUL RETRIEVAL TASK

Feature Extractor	lambda	P@1	P@3000	P@8000	mAP
Resnet18	λ=0.5	0.6456	0.6068	0.5747	0.5921
	λ=0.05	0.7154	0.7258	0.6746	0.6948
	λ=0.005	0.8109	0.7769	0.7001	0.7696
	λ=0.0005	0.9920	0.9813	0.9872	0.9898
	λ=0.00005	0.9786	0.9734	0.9758	0.9762
Resnet50	λ=0.5	0.6618	0.6190	0.5918	0.6029
	λ=0.05	0.7319	0.7469	0.6913	0.7189
	λ=0.005	0.8210	0.7913	0.7321	0.7814
	λ=0.0005	0.9954	0.9901	0.9958	0.9917
	λ=0.00005	0.9803	0.9832	0.9799	0.9801

TABLE VII
QUANTIFYING THE EFFECTIVE OF DIFFERENT LAMBDA ON PAN RETRIEVAL TASK

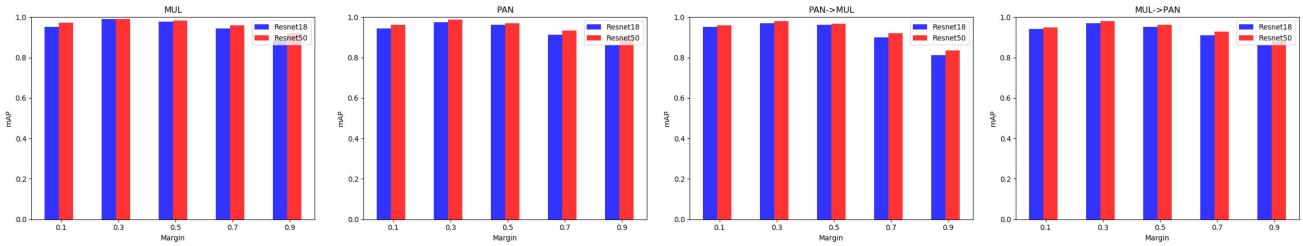
Feature Extractor	lambda	P@1	P@3000	P@8000	mAP
Resnet18	λ=0.5	0.6402	0.5766	0.5002	0.5649
	λ=0.05	0.7016	0.7365	0.6299	0.6731
	λ=0.005	0.9813	0.9725	0.9802	0.9762
	λ=0.0005	0.9532	0.9508	0.9628	0.9545
	λ=0.00005	0.9387	0.9412	0.9456	0.9398
Resnet50	λ=0.5	0.6549	0.5813	0.5210	0.5713
	λ=0.05	0.7135	0.7428	0.6456	0.6987
	λ=0.005	0.9802	0.9903	0.9810	0.9892
	λ=0.0005	0.9825	0.9834	0.9769	0.9815
	λ=0.00005	0.9707	0.9799	0.9734	0.9786

TABLE VIII
QUANTIFYING THE EFFECTIVE OF DIFFERENT LAMBDA ON MUL ->PAN RETRIEVAL TASK

Feature Extractor	lambda	P@1	P@3000	P@8000	mAP
Resnet18	λ=0.5	0.6234	0.5825	0.5467	0.5658
	λ=0.05	0.7856	0.7023	0.6456	0.6692
	λ=0.005	0.7789	0.7532	0.7692	0.7669
	λ=0.0005	0.9897	0.9698	0.9786	0.9701
	λ=0.00005	0.9578	0.9412	0.9489	0.9598
Resnet50	λ=0.5	0.6487	0.6023	0.5758	0.5902
	λ=0.05	0.7102	0.7268	0.6757	0.7021
	λ=0.005	0.8021	0.7534	0.7213	0.7611
	λ=0.0005	0.9898	0.9798	0.9714	0.9811
	λ=0.00005	0.9769	0.9821	0.9349	0.9719

TABLE IX
QUANTIFYING THE EFFECTIVE OF DIFFERENT LAMBDA ON PAN->MUL RETRIEVAL TASK

Feature Extractor	lambda	P@1	P@3000	P@8000	mAP
Resnet18	λ=0.5	0.6315	0.5514	0.4797	0.5218
	λ=0.05	0.6845	0.7129	0.6108	0.6431
	λ=0.005	0.9769	0.9658	0.9601	0.9679
	λ=0.0005	0.9421	0.9316	0.9443	0.9417
	λ=0.00005	0.9290	0.9208	0.9213	0.9202
Resnet50	λ=0.5	0.6421	0.5523	0.5108	0.5416
	λ=0.05	0.7026	0.7234	0.6063	0.6762
	λ=0.005	0.9781	0.9875	0.9702	0.9798
	λ=0.0005	0.9778	0.9713	0.9782	0.9745
	λ=0.00005	0.9633	0.9625	0.9587	0.9613



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

Fig. 8. Quantifying the effects of different margin for different tasks: (a) the model with different margin on MUL retrieval task, (b) The model with different margin on PAN retrieval task, (a) the model with different margin on PAN->MUL retrieval task, (b) The model with different margin on MUL->PAN retrieval task.

TABLE X
QUANTIFYING THE EFFECTIVE OF DIFFERENT CONV. LAYER ON MUL->PAN RETRIEVAL TASK

Feature Extractor	Frozen	P@1	P@3000	P@8000	mAP
Resnet18	Conv1-Conv5	0.6218	0.6913	0.6915	0.6334
	Conv2-Conv5	0.8632	0.8215	0.8931	0.8692
	Conv3-Conv5	0.9897	0.9698	0.9786	0.9701
	Conv4-Conv5	0.9013	0.9402	0.9310	0.9568
	Conv5	0.7881	0.8003	0.7802	0.7987
Resnet50	Conv1-Conv5	0.7013	0.7624	0.7814	0.7132
	Conv2-Conv5	0.9231	0.9019	0.9513	0.9421
	Conv3-Conv5	0.9898	0.9798	0.9714	0.9811
	Conv4-Conv5	0.9323	0.9443	0.9665	0.9611
	Conv5	0.8469	0.8021	0.8210	0.8243

layer on the cross-source retrieval tasks. The performance of this method is very sensitive with the frozen layer from the study's results. Freezing the Conv1-Conv5 layers or Conv2-Conv5 layers, makes learning of some rich features on another new sources difficult. Furthermore, freezing the Conv4-Conv5 layers or Conv5 layers, makes the retaining of the high-level semantic features of the sources that have already been trained difficult. However, freezing of Conv3-Conv5 layer gives optimal results.

To verify the robustness of the proposed model for different similarity measures, results of two features with four similarity indicators in [50] are presented. The Euclidean distance Q_E, mean absolute error (MAE) Q_{MAE}, cosine coefficients $Q_{\cos}$ and correlation coefficient (CC) Q_{CC} are introduced. In Table XII,

TABLE XI
QUANTIFYING THE EFFECTIVE OF DIFFERENT CONV. LAYER ON PAN->MUL RETRIEVAL TASK

Feature Extractor	Frozen	P@1	P@3000	P@8000	mAP
Resnet18	Conv1-Conv5	0.6913	0.6216	0.6021	0.6122
	Conv2-Conv5	0.8732	0.8205	0.8813	0.8152
	Conv3-Conv5	0.9769	0.9658	0.9601	0.9679
	Conv4-Conv5	0.9678	0.9721	0.9032	0.9445
	Conv5	0.8413	0.8210	0.7971	0.8111
Resnet50	Conv1-Conv5	0.6675	0.6841	0.6810	0.6921
	Conv2-Conv5	0.8196	0.8632	0.8332	0.8546
	Conv3-Conv5	0.9781	0.9875	0.9702	0.9798
	Conv4-Conv5	0.9210	0.9515	0.9616	0.9412
	Conv5	0.8032	0.8321	0.8018	0.8213

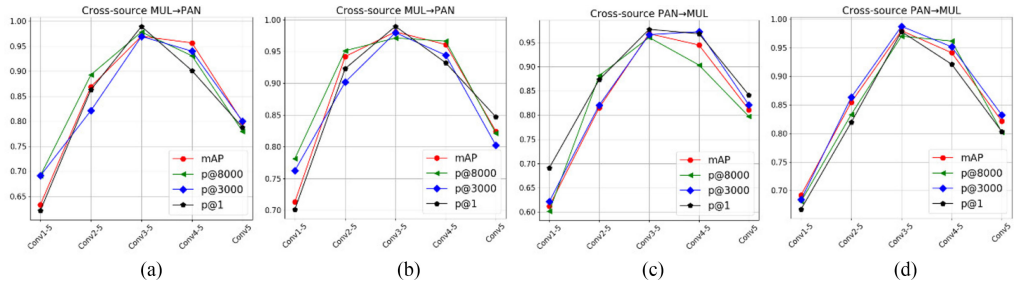


Fig. 9. Comparisons in mAP, P@1, P@3000 and P@8000 for different tasks. (a) Comparisons with different Conv. Layer on MUL→PAN retrieval task. (b) Comparisons with different Conv. Layer on MUL→PAN retrieval task. (c) Comparisons with different Conv. Layer on PAN→MUL retrieval task. (d) Comparisons with different Conv. Layer on PAN→MUL retrieval task.

TABLE XII
MAP VALUES OF TWO FEATURES UNDER DIFFERENT INDICATORS

Feature	Indicators	MUL	PAN	MUL-> PAN	PAN -> MUL
Ft	Q_E	0.9702	0.9631	0.9608	0.9601
	Q_{MAE}	0.9712	0.9645	0.9611	0.9606
	Q_{cos}	0.9903	0.9854	0.9809	0.9801
	Q_{CC}	0.9813	0.9779	0.9782	0.9715
Fi	Q_E	0.9917	0.9892	0.9811	0.9798
	Q_{MAE}	0.9909	0.9884	0.9806	0.9794
	Q_{cos}	0.9921	0.9898	0.9817	0.9802
	Q_{CC}	0.9915	0.9886	0.9807	0.9795

three losses constrain the same feature Ft, while Fi is only constrained by softmax loss by adding the BN layer. It can be seen that the results are imbalanced for Ft, and $Q_{\cos}$ yields better results than other indicators. Notably, the performance of Fi is comparable and close. This is mainly because different goals of softmax loss and triplet loss limit the optimization of this model, and the constraint is largely reduced by addition of the BN layer.

E. Comparison With Several Baselines

By comparing the optimal results of the proposed method with SIDHCNNs and several baselines from [5], we demonstrate its superiority. The experiments are conducted under the same experimental setups as SIDHCNNs. With the proposed distillation network, any architecture network can be chosen as the feature extractor. In this study, the shallower network, Resnet18, and a deeper network, resnet50 are adopted and combined with proposed distillation network to solve the CS-CBRSIR task and compared with baselines. The hyperparameter λ is set to 0.0005 on MUL→PAN retrieval task and λ is set to 0.005 on PAN→MUL retrieval task, and the freezing layer to Conv3-Conv5 as discussed above.

We summarize the mAP values of various methods in Table XIII. Distillation_Res18 denotes the proposed method with the architecture network Resnet18, and Distillation_Res50 denotes the method with the architecture network Resnet50.

TABLE XIII
COMPARISON OF MAP UNDER DIFFERENT METHODS

Methods	PAN→MUL	MUL→PAN
CCA [45]	0.1502	0.1505
SCM [46]	0.3767	0.3871
DVAN[25]	0.7162	0.7195
One-Stream[48]	0.7812	0.7903
Two-Stream[48]	0.7645	0.7682
Zero-Padding[48]	0.8031	0.8065
TONE[49]	0.7823	0.7894
TONE+HCML[49]	0.8216	0.8301
SIDHCNNs(l=8) [5]	0.9473	0.9668
SIDHCNNs(l=16) [5]	0.9552	0.9725
SIDHCNNs(l=32) [5]	0.9643	0.9789
Distillation_Res18	**0.9697**	0.9701
Distillation_Res50	**0.9798**	**0.9811**

TABLE XIV
COMPARISON OF MAP UNDER DIFFERENT INDICATORS

Feature	Indicators	MUL-> PAN	PAN -> MUL
SIDHCNNs (l=32) [5]	Q_E	0.9789	0.9643
	Q_{MAE}	0.9603	0.9582
	Q_{cos}	0.9018	0.8973
	Q_{CC}	0.9102	0.9087
Distillation _Res50	Q_E	0.9811	0.9798
	Q_{MAE}	0.9806	0.9794
	Q_{cos}	0.9817	0.9802
	Q_{CC}	0.9807	0.9795

The different l in method SIDHCNNs denotes the length of hashing feature code. The table list, the CCA [45] and SCM [46] adopt the handcrafted features, while other methods adopt deep features. From the comparisons, the handcrafted features result into unsatisfactory performance in both of the two cross-source retrieval tasks. This is because the representation ability of handcrafted feature cannot effectively represent the rich semantic information on CS-CBRSIR task. Notably, the DAVN [25] significantly outperforms the approaches based on handcrafted features. However, the DAVN provides a solution for the retrieval of spoken audio and remote sensing image, which is not effective enough to solve the problem of CS-CBRSIR. The architecture of feature learning method (one-stream [48], two-stream [48] and zero-padding [48]) and metric learning method (TONE [49] and TONE+HCML [49]) has the ability to capture the source-specific information; but the purpose of these methods is to solve cross-model problem in nature scene. The complexity of CS-CBRSIR task limits the effectiveness of these methods. Compared to the methods in [5] the proposed Distillation_Res18 yields comparable or even better results on PAN→MUL retrieval task. The Distillation_Res50 described in this study achieve optimal retrieval performance on both PAN→MUL and MUL→PAN retrieval tasks.

Besides, as one of the core issues in CBRSIR, similarity measure influences the performance of retrieval. We compare our method with the best results of the method in [5] with different similarity indicators in [50]. As shown in Table XIV, when Q_E and Q_{MAE} indicators are compared with $Q_{\cos}$ and Q_{CC} indicators, the mAP values of SIDHCNNs ($l = 32$) [5] decrease significantly for $Q_{\cos}$ and Q_{CC} indicators. This is mainly because four optimization constraints are designed in the final fully convolutional layer in the SIDHCNNs ($l = 32$), which only optimizes the Euclidean distances, but does not address the cosine distances in the feature space. To obtain more discriminative features, three losses are introduced in our method to constrain two different feature layers separately with a BN layer. The results obtained are close and stable demonstrating the robustness of our method.

V. CONCLUSION

In this article, we have proposed a novel cross-source distillation network for CS-CBRSIR. To our knowledge, this is the first work to solve the problem of CS-CBRSIR using knowledge distillation. In the proposed method, the feature extractor with the JOC, is employed to extract rich semantic information from the first sources. Afterward, the learned features of the first source are dedicated to the second source. In the training of the second source, all the midlevel to high-level layers are frozen, which retains the high-level semantic features from the first source and learns important low-level features from the second source. The performance of the proposed method is tested through a detailed ablation study was conducted on the only public dataset DSRSID. The results demonstrate that the performance proposed method is satisfactory.

Table XIII shows that the proposed method achieves high value on CS-CBRSIR task. This method requires further improvement given that its performance is tested on the only public dataset DSRSID which is relatively simple. Future works should explore the performance of this method on more challenging dataset.

REFERENCES

[1] M. Chi, A. Plaza, J. A. Benediktsson, Z. Sun, J. Shen, and Y. Zhu, "Big data for remote sensing: Challenges and opportunities," *Proc. IEEE*, vol. 104, no. 11, pp. 2207–2219, Nov. 2016.
[2] L. Wang, H. Zhong, R. Ranjan, A. Zomaya, and P. Liu, "Estimating the statistical characteristics of remote sensing big data in the wavelet transform domain," *IEEE Trans. Emerg. Topics Comput.*, vol. 2, no. 3, pp. 324–337, Sep. 2014.
[3] P. Du, Y. Chen, T. Hong, and F. Tao, "Study on content-based remote sensing image retrieval," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2005, pp. 707–710.
[4] D. Li, "Content-based remote sensing image retrieval," *Proc. SPIE*, vol. 6044, 2005, Art. no. 60440Q.
[5] Y. Li, Y. Zhang, and X. Huang, "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.

[6] G. Healey and A. Jain, "Retrieving multispectral satellite images using physics-based invariant representations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 8, pp. 842–848, Aug. 1996.

[7] E. Aptoula, "Remote sensing image retrieval with global morphological texture descriptors," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 3023–3034, May 2014.

[8] G. J. Scott, M. N. Klaric, C. Davis, and H. C. Shyu, "Entropy-balanced bitmap tree for shape-based object retrieval from large-scale satellite imagery databases," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 5, pp. 1603–1616, May 2011.

[9] D.G. Lowe, "Object recognition from local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, Sep. 1999, pp. 1150–1157.

[10] H. Bay, T. Tuytelaars, and L.V. Gool, "Surf: Speeded up robust features," in *Proc. Eur. Conf. Comput. Vis.*, May 2006, pp. 404–417.

[11] J. Yang, J. Liu, and Q. Dai, "An improved bag-of-words framework for remote sensing image retrieval in large-scale image databases," *Int. J. Digit. Earth*, vol. 8, no. 8, pp. 273–292, 2015.

[12] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3304–3311.

[13] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[14] T. Jiang, G. Xia, Q. Lu, and W. Shen, "Retrieving aerial scene images with learned deep image-sketch features," *J. Comput. Sci. Technol.*, vol. 32, no. 8, pp. 726–737, 2017

[15] W. Zhou, S. Newsam, C. Li, and Z. Shao, "Learning low dimensional convolutional neural networks for high-resolution remote sensing image retrieval," *Remote Sens.*, vol. 9, no. 8, p. 489, 2017.

[16] R. Cao, Q. Zhang, and J. S. Zhu, "Enhancing remote sensing image retrieval with triplet deep metric learning network," *Int. J. Remote Sens.*, 2019, *arXiv:1902.05818*.

[17] W. Zhou, S. Newsam, C. Li, and Z. Shao, "Learning low dimensional convolutional neural networks for high-resolution remote sensing image retrieval," *Remote Sens.*, vol. 9, no. 4, p. 489, 2017.

[18] Y. Ge, S. Jiang, Q. Xu, C. Jiang, and F. Ye, "Exploiting representations from pre-trained convolutional neural networks for high-resolution remote sensing image retrieval," *Multimedia Tools Appl.*, pp. 1–27, 2017.

[19] W. Zhou, S. Newsam, C. Li, and Z. Shao, "Learning low dimensional convolutional neural networks for high-resolution remote sensing image retrieval," *Remote Sens.*, vol. 9, no. 3, p. 489, 2016.

[20] R. Imbriaco, C. Sebastian, E. Bondarev, and H. Peter, "Aggregated deep local features for remote sensing image retrieval," *Remote Sens.*, vol. 11, no. 8, p. 5, 2019.

[21] P. Li, P. Ren, X. Zhang, Q. Wang, X. Zhu, and L. Wang, "Region-wise deep feature representation for remote sensing images," *Remote Sens.*, vol. 10, no. 6, p. 871, 2018.

[22] 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. 4, p. 281, 2019.

[23] Y. Li, Y. Zhang, X. Huang, H. Zhu, and J. Ma, "Large-scale remote sensing image retrieval by deep hashing neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 950–965, Feb. 2018.

[24] B. Chaudhuri, B. Demir, S. Chaudhuri, and L. Bruzzone, "Multilabel remote sensing image retrieval using a semisupervised graph-theoretic method," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 1144–1158, Feb. 2017.

[25] M. Gou, Y. Yuan, and X. Lu, "Deep cross-modal retrieval for remote sensing image and audio," in *Proc. IEEE 10th Int. Assoc. Pattern Recognit. Workshop Pattern Recognit. Remote Sens.*, 2018, pp. 1–7.

[26] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2183–2195, Apr. 2018.

[27] Q. Y. Jiang and W. J. Li, "Deep cross-modal hashing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3270–3278.

[28] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *Proc. Neural Inf. Process. Syst. Deep Learn. Representation Learn. Workshop*, *Preprints*, 2015, *arXiv:1503.02531*.

[29] G. Saurabh, H. Judy, and M. Jitendra, "Cross modal distillation for supervision transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2827–2836.

[30] Z. Sergey and K. Nikos, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," in *Proc. 8th Int. Conf. Learn. Representations*, *Preprints*, 2016, *arXiv:1612.03928*.

[31] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7130–7138.

[32] Y. T. Chen, N. Y. Wang, and Z. X. Zhang, "DarkRank: Accelerating deep metric learning via cross sample similarities transfer," in *Proc. Assoc. Advancement Artif. Intell. Conf. Artif. Intell.*, *Preprints*, 2018. *arXiv:1707.01220*.

[33] Z. Xu, Y. C. Hsu, and J. W. Huang, "Training shallow and thin networks for acceleration via knowledge distillation with conditional adversarial networks," *Preprints*, 2017, *arXiv:1709.00513*.

[34] Z. Shao, W. Zhou, L. Zhang, and J. Hou, "Improved color texture descriptors for remote sensing image retrieval." *J. Appl. Remote Sens.* vol. 8, no. 6, 2014, Art. no. 83584.

[35] X. Zhu and Z. Shao, "Using no-parameter statistic features for texture image retrieval," *Sens. Rev.*, vol. 31, no. 8, pp. 144–153, 2011

[36] X. Jin, F. Yi, and Z. Fan, "Deepshape: Deep learned shape descriptor for 3D shape matching and retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1275–1283.

[37] W. Zhou, X. Deng, and Z. Shao, "Region convolutional features for multi-label remote sensing image retrieval," 2018, *arXiv:1807.08634*.

[38] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[39] M. Wang and T. Song, "Remote sensing image retrieval by scene semantic matching," *IEEE Trans. Geosci. Remote Sens.* vol. 51, no. 3, pp. 2874–2886, May 2013.

[40] O. A. Penatti, K. Nogueira, and J. A. Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2015, pp. 44–51.

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

[42] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. 14th Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 499–515.

[43] L. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 8, pp. 2579–2605, 2008

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

[45] Y. Gong, S. Lazebnik, A. Gordo, and F. Perronnin, "Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2916–2929, Dec. 2013.

[46] D. Zhang and W. Li, "Large-scale supervised multimodal hashing with semantic correlation maximization," in *Proc. 28th Assoc. Advancement Artif. Intell. Conf. Artif. Intell.*, 2014, pp. 2177–2183.

[47] Q. Liu, R. Hang, H. Song, and Z. Li, "Learning multiscale deep features for high-resolution satellite image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 1, pp. 117–126, Jan. 2018.

[48] A. Wu, W. Zheng, H. Yu, S. Gong, and J. Lai, "RGB-infrared cross-modality person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5390–5399.

[49] M. Ye, X. Lan, J. Li, and P. C. Yuen, "Hierarchical discriminative learning for visible thermal person re-identification," in *Proc. 32nd Assoc. Advancement Artif. Intell. Conf. Artif. Intell.*, 2018.

[50] X. Li, H. Shen, H. Li, and L. Zhang, "Patch matching-based multitemporal group sparse representation for the missing information reconstruction of remote-sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 8, pp. 3629–3641, Aug. 2016.

[51] Q. Liu, R. Hang, H. Song, and F. Zhu, "Adaptive deep pyramid matching for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, Nov. 2016.

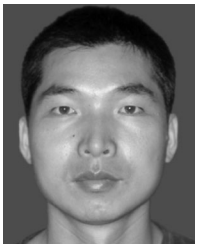
[52] Q. Liu, F. Zhou, R. Hang, and X. Yuan, "Bidirectional-convolutional LSTM based spectral-spatial feature learning for hyperspectral image classification," *Remote Sens.*, vol. 9, no. 12, 2017, Art. no. 1330.

[53] S. M. M. Kahaki, A. Haslina, N. M. Jan, and I. Waidah, "Geometric feature descriptor and dissimilarity-based registration of remotely sensed imagery," *PLoS ONE*, vol. 13, no. 7, 2018.

[54] S. M. M. Kahaki, M. J. Nordin, A. H. Ashtari, and S. J. Zahra, "Invariant feature matching for image registration application based on new dissimilarity of spatial features," *PLoS ONE*, vol. 11, no. 3, 2016.

[55] Q. Yuan, Y. Wei, X. Meng, H. Shen, and L. Zhang, "A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 3, pp. 978–989, Mar. 2018.

[56] Z. Li and H. Shen, "Cloud detection by fusing multiscale convolutional features," ISPRS Technical Commission III on Remote Sensing. Proceedings of the ISPRS Technical Commission III Midterm Symposium on "Developments,Technologies and Applications in Remote Sensing, 2018, pp. 137–140.



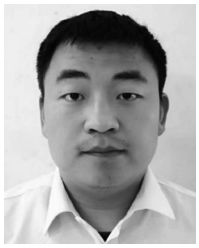
Wei Xiong received the B.S., M.S., and Ph.D. degrees from Naval Aviation University, Yantai, China, in 1998, 2001, and 2005, respectively.

From 2007 to 2009, he was a Postdoctoral Researcher with the Department of Electronic Information Engineering, Tsinghua University, Beijing, China. He is currently a Full Professor with the Naval Aviation University, Yantai, China, where he teaches Random Signal Processing and Information Fusion. He is one of the Founders and the Directors of Research Institute of Information Fusion, Naval Aviation University. His research interests include pattern recognition, remote sensing, and multisensor information fusion.

Dr. Xiong is the Member and Director General of Information Fusion Branch of Chinese Society of Aeronautics and Astronautics.

Zhenyu Xiong received the B.S. degrees from Naval Aviation University, Yantai, China, in 2014 and 2018, respectively, and is currently working toward the M.S. degree in information and communication engineering with Naval Aviation University.

His research interests include information fusion and deep learning with their applications in remote sensing.



Yaqi Cui received the B.S., M.S., and Ph.D. degrees in information and communication engineering from Naval Aviation University, Yantai, China, in 2008, 2011, and 2014, respectively.

Since 2014, he has been a Lecturer with Naval Aviation University. His research interests include information fusion, machine learning, and deep learning with their applications in information fusion.

Yafei Lv received the B.S. and M.S. degrees from Military Transportation University, Tianjin, China, in 2014 and 2017, respectively, and is currently working toward the Ph.D. degree in information and communication engineering with Naval Aviation University. His research interests include computer vision and deep learning with their applications in remote sensing.