remote sensing

MDPI

Article

Dual Modality Collaborative Learning for Cross-Source Remote Sensing Retrieval

Jingjing Ma, Duanpeng Shi, Xu Tang *, Xiangrong Zhang and Licheng Jiao

School of Artificial Intelligence, Xidian University, Xi'an 710071, China; jjma@xidian.edu.cn (J.M.); 19171213861@stu.xidian.edu.cn (D.S.); xrzhang@mail.xidian.edu.cn (X.Z.); lchjiao@mail.xidian.edu.cn (L.J.)
* Correspondence: tangxu128@xidian.edu.cn

Abstract: Content-based remote sensing (RS) image retrieval (CBRSIR) is a critical way to organize high-resolution RS (HRRS) images in the current big data era. The increasing volume of HRRS images from different satellites and sensors leads to more attention to the cross-source CSRSIR (CS-CBRSIR) problem. Due to the data drift, one crucial problem in CS-CBRSIR is the modality discrepancy. Most existing methods focus on finding a common feature space for various HRRS images to address this issue. In this space, their similarity relations can be measured directly to obtain the cross-source retrieval results straight. This way is feasible and reasonable, however, the specific information corresponding to HRRS images from different sources is always ignored, limiting retrieval performance. To overcome this limitation, we develop a new model for CS-CBRSIR in this paper named dual modality collaborative learning (DMCL). To fully explore the specific information from diverse HRRS images, DMCL first introduces ResNet50 as the feature extractor. Then, a common space mutual learning module is developed to map the specific features into a common space. Here, the modality discrepancy is reduced from the aspects of features and their distributions. Finally, to supplement the specific knowledge to the common features, we develop modality transformation and the dual-modality feature learning modules. Their function is to transmit the specific knowledge from different sources mutually and fuse the specific and common features adaptively. The comprehensive experiments are conducted on a public dataset. Compared with many existing methods, the behavior of our DMCL is stronger. These encouraging results for a public dataset indicate that the proposed DMCL is useful in CS-CBRSIR tasks.

Keywords: cross-source content-based remote sensing image retrieval; high-resolution remote sensing; modality discrepancy

Citation: Ma, J.; Shi, D.; Tang, X.; Zhang, X.; Jiao, L. Dual Modality Collaborative Learning for Cross-Source Remote Sensing Retrieval. *Remote Sens.* **2022**, *14*, 1319. https://doi.org/10.3390/rs14061319

Academic Editor: Pedro Melo-Pinto

Received: 11 February 2022
Accepted: 8 March 2022
Published: 9 March 2022

1. Introduction

With the advancement in remote sensing (RS) observation technologies, the capability of capturing RS images has been enhanced dramatically. An enormous volume and a large variety of high-resolution RS (HRRS) images, therefore, can be collected every day. The HRRS image processing has entered the big data era [1–4]. To obtain valuable information from these HRRS images, the first step is to manage them reasonably and intelligently according to users' opinions. Therefore, content-based remote sensing retrieval (CBRSIR) attracts researchers' attention. As a useful image management tool, CBRSIR plays an essential role in broad applications, such as land cover extraction, energy optimization, and agriculture and forest monitoring [5–7].

In recent decades, many useful methods have been developed for unified-source CBRSIR (US-CBRSIR) tasks [8]. In US-CBRSIR, the query and the target images (within the image archive) are from the same RS source. For example, both of them are Synthetic Aperture Radar (SAR) images [9,10] or high-resolution optical RS images [11,12]. Feature extraction/learning is of vital importance for US-CBRSIR. Numerous feature descriptors,

ranging from hand-crafted features [13] to deep-learning-based features [14,15], are exploited and applied to map the HRRS images into discriminative features. Then, simple or specific distance metrics [16] are designed to complete US-CBRSIR according to the resemblance between features.

With the volume and types of RS images increasing dramatically, scholars need to process diverse HRRS images collected by various sensors. In this scenario, the query and the target images may be from different RS sources. Thus, cross-source CBRSIR (CS-CBRSIR) is proposed. CS-CBRSIR can be seen as a member of the cross-modal family [17–21], and it is confronted with the challenge of heterogeneity gaps [22] when measuring the resemblance between different types of HRRS data. Another challenge in CS-CBRSIR is the data shift problem [23] where the data distributions are different as the source and target images are acquired with various sensors. Therefore, the aforementioned feature descriptor-based methods, which are widely used in US-CBRSIR, cannot extend to CS-CBRSIR directly since they do not consider the challenge discussed above.

Although many useful cross-modal retrieval methods have been introduced for natural images and they achieve successes in different applications, we cannot apply them to deal with HRRS CS-CBRSIR tasks immediately. The reasons can be summarized as follows. First, the feature extraction blocks in the natural cross-modal methods do not take the characteristics of HRRS images into account. The extracted features are not fully capable of describing the complex contents and intricate structures of HRRS images. Second, many existing methods (such as [24]) reduce the influence of modality discrepancy through the mono-directional knowledge transferring, i.e., they transfer the knowledge from one source to another. However, the mutual effect of images from different sources is not considered. Third, those methods emphasize the importance of common space learning but ignore the modality-specific features, which are also crucial to the CS-CBRSIR retrieval task.

To overcome the limitations mentioned above, we propose a new network for CS-CBRSIR under the cross-modal paradigm considering properties of different HRRS images. We name it dual modality collaborative learning (DMCL). First, a two-stream network is developed to extract specific features from HRRS images. Second, a common space mutual learning module is introduced to project the obtained specific features into a common space. Third, the modality transform scheme is designed to map specific features from one source to the other source mutually. Finally, the obtained common and specific features are further fused to endow the robustness of the final representation in cross-source retrieval tasks.

The main contributions of this paper are summarized as follows:

- A new HRRS CS-CBRSIR method (DMCL) is proposed based on the framework of cross-modal learning in this paper. DMCL can learn the discriminative specific and common features from different types of HRRS images, which are beneficial to CS-CBRSIR tasks.
- A common mutual learning module is developed to eliminate modality discrepancy, in which the information corresponding to different sources is forced to exchange reciprocally. Thus, the influence of modality discrepancy can be reduced to the greatest extent.
- The developed dual-space feature fusion module with the modality transform scheme ensures that the HRRS images from different sources can be represented comprehensively. Thus, the distances obtained by those representations can reflect the valid similarity relationships between different RS images.

The rest of this paper is organized as follows. The literature related to US-CBRSIR and CS-CBRSIR is reviewed in Section 2. In Section 3, our model is introduced in detail. The experiments and the discussion are reported in Section 4. Section 5 provides a brief conclusion.

2. Related Work

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

In general, a regular CBIR system includes two modules [25]. One is the feature extraction module which can draw effective features to characterize both the input query and target images. The other is the retrieval module that returns a ranked list of similar images to a query image. As an application of CBIR, the essential components of US-CBRSIR are the same as the modules discussed above. In this section, we review some existing US-CBRSIR methods. For clarity, we group them into two sets, including methods based on hand-crafted features and approaches based on deep features.

The hand-crafted visual feature descriptors can be divided into two categories, i.e., low- and mid-level descriptors. The popular low-level features contain spectral [26,27], texture [28–30], shape [31,32], etc. These descriptors are extracted from RS images based on engineering skills and domain expertise. The authors of [33] mine the contents of RS images by exploring spectral distribution information. RS images are first segmented into different regions. Then, the spectral information corresponding to diverse regions is extracted and aggregated to describe RS images for US-CBRSIR. Shao et al. [34] present two feature descriptors, including Gabor wavelet texture (CGWT) and color Gabor opponent texture (CGOT), to exploit discriminative information from RS images. Those two features can describe the diverse objects within RS images. Although low-level features are easy to accomplish and stable in performance, the shallow representational ability limits their performance in retrieval tasks. To solve this problem, mid-level features appear and achieve success in US-CBRSIR. The prevalent mid-level features include bag-of-words (BOW) [35], Fisher vectors [36], and vector of locally aggregated descriptors (VLAD) [37]. Compared with low-level features, they are capable of representing the complex context hidden in RS images [38]. For example, morphological texture descriptors are combined with the bag-of-words paradigm in [39] to deeply mine the contents of RS images. The global morphological descriptors are obtained using subwindows, and they are then clustered to form a vocabulary of visual words. The contents of image are represented by the frequency histogram of the visual words. In [40], a local pattern spectrum is employed as a morphological descriptor, which is calculated from a whole image with a dense grid with fixed steps. It is then combined with VLAD to generate a visual vocabulary. The extracted image features from the above methods are low-dimensional, which would speed up the retrieval speed. The mentioned methods based on the hand-crafted features have achieved their successes in US-CBRSIR. However, they cannot reach the satisfactory stage due to the well-known semantic gap [41] which occurs between the low-level features and high-level semantics.

In recent years, with the development of deep learning [42,43], especially the deep convolutional neural network (DCNN), the high-level features play significant roles in numerous RS applications as well as US-CBRSIR. Due to the specific data-driven learning scheme, complex contents within HRRS images can be fully represented by high-level features. Many deep-based US-CBRSIR methods have been proposed [8]. Zhou et al. [44] develop a shallow CNN to learn the information from RS images. It consists of five convolution layers, a global average pooling layer, and an mlp-conv layer. The reported positive retrieval results confirm the usefulness of deep features in US-CBRSIR. To get more robust and discriminative deep features for US-CBRSIR, some large CNN models are introduced. Nevertheless, these models need a lot of labeled data to complete the training process, which is a harsh or even impractical condition for RS images. Therefore, transfer learning attracts scholars' attention. Instead of training the CNNs directly, researchers use a limited amlount of labeled RS data to fine-tune the pre-trained CNNs (such as AlexNet [45]) for learning the deep features from RS images. At the same time, some unsupervised deep feature learning methods have been introduced to tackle US-CBRSIR. For instance, under the framework of BOW, an unsupervised deep feature learning method is proposed in [14] based on a convolutional auto-encoder. Rather than learning the features from RS images directly, the authors mine knowledge at the patch level. Then, varied RS

patch representation is fused by the codebook mapping. Deep features perform well in US-CBRSIR. However, they are always dense and high-dimensional, which limits the retrieval efficiency. To overcome this limitation, hashing techniques enter the US-CBRSIR field. By mapping continual features into discrete binary hash codes, the dimensions of features can be reduced, and the retrieval speed can be increased. A semi-supervised deep hashing network is developed in [11]. In this method, RS images are first mapped into deep features through DCNNs. Then, hash codes are obtained by the adversarial auto-encoder [46] with the specific binary constraints. Liu et al. [12] proposes a deep hashing network to learn the discrete and compact hash codes from RS images. Due to the strong capacity of feature learning and the simple Hamming distance metric, the US-CBRSIR method based on binary hash codes has good performance and is highly efficient in terms of speed, but at the cost of accuracy.

2.2. Cross-Modal Retrieval in Remote Sensing

In the RS community, the cross-modal retrieval scenario mainly includes the audio–visual, text–visual, and visual–visual retrieval tasks [47]. Guo et al. [48] propose the deep audio–visual network (DVAN) and constructed a remote sensing audio–image caption dataset, which provides a new method for RS image retrieval. As for the text–visual retrieval task, Lu et al. [49] release a remote sensing image captioning dataset (RSICD) and introduce a normal network that contains a CNN and RNN or LSTM to predict words for RS images. Yuan et al. [50] use multilevel attention to focus on features of both specific spatial and multiple scales. The attribute graph is employed to learn more useful attribute features for image captioning.

Our work focuses on the cross-source image–image retrieval task. The first CS-CBRSIR work might have been proposed in [51], in which a dual-source remote sensing dataset (DSRSID) is released, and a dual-hashing net is developed to learn features from RS images collected by different sensors. Specifically, DSRSID contains two different types of RS images, i.e., panchromatic and multispectral images. Considering their properties, a DCNN is developed to map RS images corresponding to various sources into binary codes for CS-CBRSIR. The reported retrieval results illustrate its usefulness. Taking the issue of common space projection into account, Li et al. [52] propose a two-stage learning model for CS-CBRSIR tasks. It uses the features extracted from one source as the supervision information for the other source. Based on the knowledge distillation, the information from one source can be transferred to the other to obtain common space representation so that the cross-source retrieval can be conducted. Ushasi et al. [53] develop a two-level training protocol to deal with CS-CBRSIR tasks. First, they obtain intermediate discriminative features using two classification networks for two sources. Then, the intermediate features are fed into an encoder–decoder model to construct the unified representation, which is used to complete the retrieval process. This method is testified by the image–image and voice–image tasks. Additionally, a cycle–identity–generative adversarial network (CI-GAN) is proposed in [54] to accomplish CS-CBRSIR. The critical point of CI-GAN is transforming cross-source images into unified-source images.

3. The Proposed Method

3.1. The Overview of the Framework

As said in Section 1, there are two challenges in CS-CBRSIR, i.e., the heterogeneity gap issue [22] and the data shift problem [23]. Overall, the above two challenges can be regarded as the modality discrepancy problem. To reduce its negative influence on CS-CBRSIR, we develop the DMCL model in this paper. Its architecture is shown in Figure 1, which contains a specific feature extractor, a common feature learning block, and an adaptive dual-modality fusion block. The specific feature extractor consists of two pre-trained DCNNs (with the same structure but not sharing weights) and two feature-refining modules. They aim to extract visual features from different HRRS images. The common space mutual learning module within the common feature learning block aims to map the

specific features corresponding to different RS images into the common space. The obtained common features and their distributions will be analyzed deeply. The adverse impacts of modality discrepancy can be reduced along with the specific loss functions. To improve the discrimination of final representation, the specific and common features are combined in the adaptive dual-modality fusion block, which contains the modality transformation and the dual-modality feature learning modules. The modality transformation is apt to transmit the information corresponding to different HRRS images mutually, while the dual-modality feature learning module is good at fusing diverse knowledge. In this way, the impacts of the modality discrepancy problem can be further reduced.

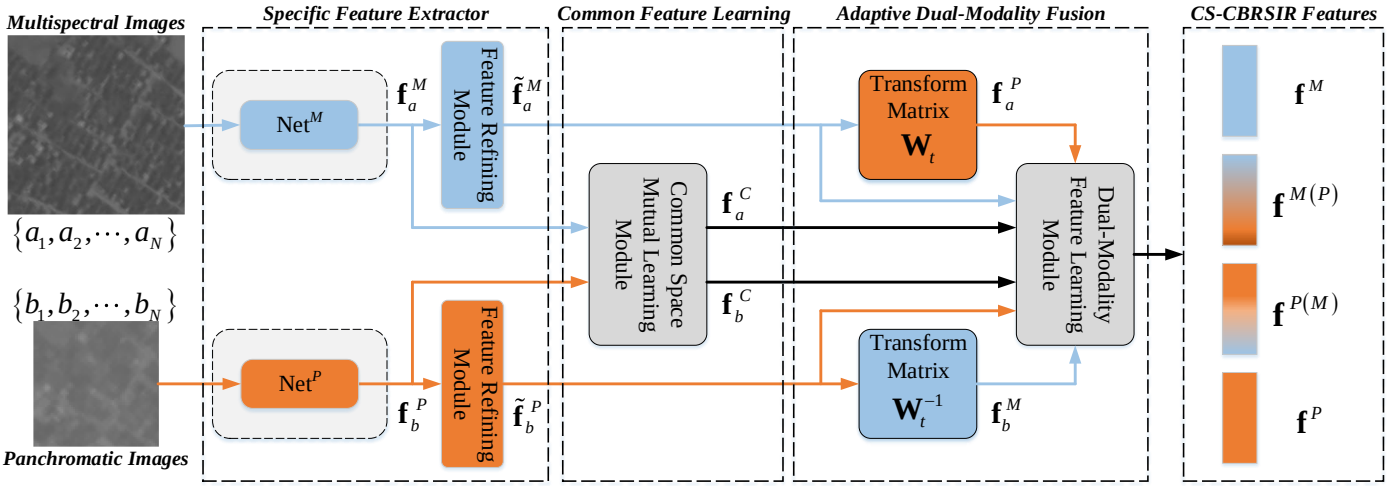

Figure 1. Framework of the proposed dual modality collaborative learning method. The multispectral and panchromatic images are first inputted into the specific feature extractor to extract their specific features $\mathbf{f}_a^M$ and $\mathbf{f}_b^P$, and refined specific features $\tilde{\mathbf{f}}_a^M$ and $\tilde{\mathbf{f}}_b^P$. Then, the common space features $\mathbf{f}_a^C$, $\mathbf{f}_b^C$ can be generated by the common feature learning block using $\mathbf{f}_a^M$ and $\mathbf{f}_b^P$. To fuse the information corresponding to different RS images and integrate the specific and common characteristics, $\tilde{\mathbf{f}}_a^M$, $\tilde{\mathbf{f}}_b^P$, $\mathbf{f}_a^C$, and $\mathbf{f}_b^C$ are inputted in the adaptive dual-modality fusion block. Then, the fused features $\mathbf{f}^M$, $\mathbf{f}^{M(P)}$, $\mathbf{f}^{P(M)}$, and $\mathbf{f}^P$ can be generated for the retrieval. The blue and orange streams illustrate the learning process of multispectral and panchromatic images, respectively.

Before introducing DMCL in detail, some preliminaries are explained here. Suppose we have a cross-source dataset consisting of multispectral (MSP) and panchromatic (PAN) image pairs $\mathcal{X} = \{(a_i, b_i, y_i)|i = 1, 2, \cdots, N\}$, where a_i and b_i denote the i-th MSP and PAN images, and y_i indicates their semantic labels. In other words, we have a set of MSP images $\mathcal{A} = \{a_1, \cdots, a_N\}$ and a set of PAN images $\mathcal{B} = \{b_1, \cdots, b_N\}$. Their semantic label set is $\mathcal{Y} = \{y_1, \cdots, y_N\}$, where $y_i \in [1, \cdots, C]$ and C means the number of classes. As a popular loss function in classification, cross-entropy $\mathcal{L}_c$ can divide the feature space into different subspaces, ensuring the deep features are discriminative. However, it pays more attention to inter-class separation. To deeply consider the intra-class compactness, we also introduce the triplet-loss $\mathcal{L}_{tri}$ [55] in this work, which can guarantee that the distances of samples in different classes are larger than those in the same category. The definitions of two essential loss functions are:

$$\begin{aligned} \mathcal{L}_c &= -\frac{1}{n}\sum_{i=1}^{n}\log(p(y_i|a_i), p = \frac{e^{\mathbf{W}_{y_i}^T\mathbf{f}_i + bias_{y_i}}}{\sum_{m=1}^{C} e^{\mathbf{W}_m^T\mathbf{f}_i + bias_m}}, \\ \mathcal{L}_{tri} &= \sum_{i=1}^{n}[d(\mathbf{f}_i, \mathbf{f}_i^{pos}) - d(\mathbf{f}_i, \mathbf{f}_i^{neg}) + \alpha], \end{aligned} \tag{1}$$

where n is the size of the mini-batch, p represents the possibility that an HRRS image is classified as the label y_i, $\mathbf{W}_m \in \mathbb{R}^d$ denotes the m-th column of the weights, $bias \in \mathbb{R}^C$ is the bias term, $\mathbf{f}_i$ is the i-th HRRS image's feature, C indicates the number of semantic

classes, $d(\cdot)$ means the Euclidean distance, $\mathbf{f}_i^{pos}$ is the feature of an HRRS image that is from the same class as the i-th HRRS image, $\mathbf{f}_i^{neg}$ is the feature of an HRRS image that is from a different class as the i-th HRRS image, and α is a margin parameter that ensures the similarities of the positive pairs are larger than those of the negative pairs. For clarity, we use $\text{Triplet}\left(\mathbf{f}_i, \mathbf{f}_i^{pos}, \mathbf{f}_i^{neg}, \alpha\right)$ and $\text{CE}(\mathbf{f}_i)$ to replace $\mathcal{L}_{tri}$ and $\mathcal{L}_c$ in the following.

3.2. Specific Feature Extractor

In this paper, we select ResNet50 [56] as the backbone for the specific feature extractor. The reasons for this selection are two-fold. First, due to the complex contents of HRRS images, the backbone should have strong nonlinear feature learning capacity. Second, taking the time complexity of our model, the backbone should be as light as possible. To sum up, ResNet50 just meets those demands. A large number of deep models have been developed based on ResNet50 to address various HRRS applications [57,58], and they achieve remarkable successes in their own applications. This demonstrates that the feature learning capacity of ResNet50 is strong. In addition, compared with some heavy models (such as GoogleNet [59]), training ResNet50 is not difficult.

As shown in Figure 1, both Net^M and Net^P contain the input header and first three residual layers of ResNet50 to extract specific features from different HRRS images. Note that Net^M and Net^P do not share parameters. When MSP images $\{a_1, a_2, \cdots, a_N\}$ and PAN images $\{b_1, b_2, \cdots, b_N\}$ are fed into Net^M and Net^P, we can obtain MSP-specific features $\mathbf{f}_a^M$ and PAN-specific features $\mathbf{f}_b^P$. Then, they are input into two feature-refining modules and the corresponding refined specific features $\tilde{\mathbf{f}}_a^M$ and $\tilde{\mathbf{f}}_a^P$ are outputted. The feature-refining module consists of a global average pooling layer and 1×1 convolutional layer, which can highlight the spatial information and reduce the amount of parameters. To ensure the discrimination of the obtained specific features, the following loss functions are defined:

$$\begin{aligned} \mathcal{L}^{Sp} &= \mathcal{L}_c^{Sp} + \mathcal{L}_{tri}^{Sp}, \\ \mathcal{L}_c^{Sp} &= \text{CE}\left(\tilde{\mathbf{f}}_a^M\right) + \text{CE}\left(\tilde{\mathbf{f}}_b^P\right), \\ \mathcal{L}_{tri}^{Sp} &= \text{Triplet}\left(\tilde{\mathbf{f}}_a^M, \left(\tilde{\mathbf{f}}_a^M\right)^{pos}, \left(\tilde{\mathbf{f}}_a^M\right)^{neg}, \alpha\right) \\ &+ \text{Triplet}\left(\tilde{\mathbf{f}}_b^P, \left(\tilde{\mathbf{f}}_b^P\right)^{pos}, \left(\tilde{\mathbf{f}}_b^P\right)^{neg}, \alpha\right). \end{aligned} \tag{2}$$

3.3. Common Feature Learning

To mitigate the modality discrepancy problem in CS-CBRSIR and measure the resemblance between HRRS images from different sources, we map the specific features $\mathbf{f}_a^M$ and $\mathbf{f}_b^P$ into a common space by the common space mutual learning module (see Figure 2). It consists of the last residual layer of ResNet50 and a 1×1 convolutional layer. Here, the 1×1 convolutional layer is used to reduce the dimensions of obtained features, and we use the L_2 normalization in the feature embedding layer. After this module, common features $\mathbf{f}_a^C$ and $\mathbf{f}_b^C$ can be obtained, which contain the common information of two data sources. To guarantee the discrimination of $\mathbf{f}_a^C$ and $\mathbf{f}_b^C$, we first formulate the following loss functions:

$$\begin{aligned} \mathcal{L}^{Cs} &= \mathcal{L}_c^{Cs} + \mathcal{L}_{tri}^{Cs}, \\ \mathcal{L}_c^{Cs} &= \text{CE}\left(\mathbf{f}_a^C\right) + \text{CE}\left(\mathbf{f}_b^C\right), \\ \mathcal{L}_{tri}^{Cs} &= \text{Triplet}\left(\mathbf{f}_a^C, \left(\mathbf{f}_a^C\right)^{pos}, \left(\mathbf{f}_a^C\right)^{neg}, \alpha\right) \\ &+ \text{Triplet}\left(\mathbf{f}_b^C, \left(\mathbf{f}_b^C\right)^{pos}, \left(\mathbf{f}_b^C\right)^{neg}, \alpha\right). \end{aligned} \tag{3}$$

Furthermore, we develop the cross-source mutual learning loss to regularize the knowledge from different sources that can be transferred reciprocally. Consequently, the representation capacity of common features can be enhanced. Particularly, three classifiers, θ^M, θ^C, and θ^P, are embedded on the top of common space mutual learning module. Then, the predictions $\mathbf{p}_a^M$, $\mathbf{p}_a^C$, $\mathbf{p}_b^C$, and $\mathbf{p}_b^P$ can be generated. Here, $\mathbf{p}_a^M$ is the output of $\mathbf{f}_a^C$ and θ^M, which implies the distributions of MSP images' specific features in the common space.

$\mathbf{p}_a^C$ and $\mathbf{p}_b^C$ are the products of $\mathbf{f}_a^C$, $\mathbf{f}_b^C$, and θ^C, which denote the distributions of common features corresponding to MSP and PAN images in the common space. $\mathbf{p}_b^P$ is the output of $\mathbf{f}_b^C$ and θ^P, which indicates the distributions of PAN images' specific features in the common space. To ensure the diverse information can be exchanged mutually, we narrow down the difference between $\mathbf{p}_a^M$ and $\mathbf{p}_b^P$, as well as the discrepancy between $\mathbf{p}_a^C$ and $\mathbf{p}_b^C$. Thus, the cross-source mutual learning loss is formulated as:

$$\begin{aligned} \mathcal{L}^{KL} &= \mathcal{L}_{CKL} + \mathcal{L}_{SpKL}, \\ \mathcal{L}_{CKL} &= \mathrm{KL}(p_a^C || p_b^C) + \mathrm{KL}(p_b^C || p_a^C), \\ \mathcal{L}_{SpKL} &= \mathrm{KL}(p_a^M || p_b^P) + \mathrm{KL}(p_b^P || p_a^M), \end{aligned} \tag{4}$$

where $\mathrm{KL}(\mathbf{p}_2||\mathbf{p}_1)$ means the Kullback–Leibler (KL) divergence [60] from $\mathbf{p}_1$ to $\mathbf{p}_2$. It is noted that KL divergence is asymmetry, i.e., $\mathrm{KL}(\mathbf{p}_2||\mathbf{p}_1) \neq \mathrm{KL}(\mathbf{p}_1||\mathbf{p}_2)$.

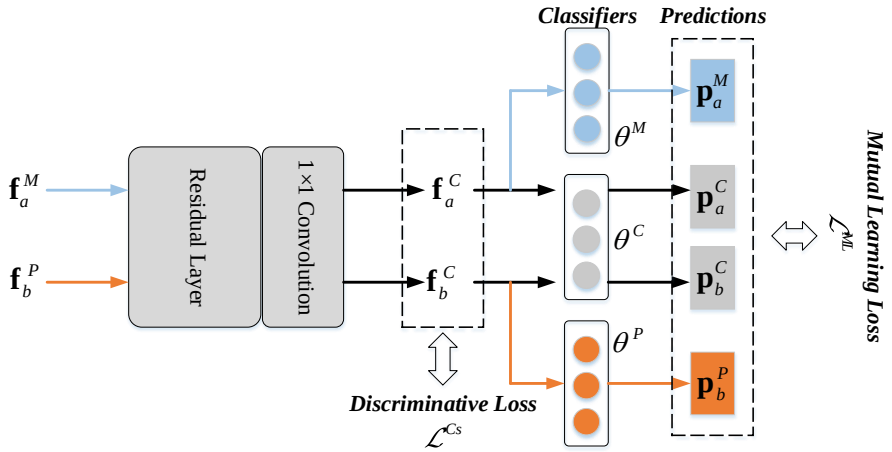

Figure 2. Framework of the common space mutual learning module.

3.4. Adaptive Dual-Modality Fusion

So far, we have obtained the representation $\mathbf{f}_a^C$ and $\mathbf{f}_b^C$ of MSP and PAN images in the common feature space. Although they can be used to accomplish CS-CBRSIR tasks, their performance would be limited due to the specific information loss. To supplement the specific knowledge, we develop the modality transformation and the dual-modality feature learning modules here.

In the modality transformation module, according to the literature [61,62], we assume that the MSP and PAN feature spaces can be transformed to each other by the linear mapping with an invertible transition matrix $\mathbf{W}_t$. Thus, two items $\tilde{\mathbf{f}}_a^M$and $\tilde{\mathbf{f}}_b^P$ corresponding to MSP and PAN spaces can be transformed into others by the following equations:

$$\begin{aligned} \mathbf{f}_a^P &= \mathbf{W}_t \cdot \tilde{\mathbf{f}}_a^M, \\ \mathbf{f}_b^M &= \mathbf{W}_t^{-1} \cdot \tilde{\mathbf{f}}_b^P, \end{aligned} \tag{5}$$

where $\mathbf{f}_a^P$ implies the representation of an MSP image in the PAN feature space, and $\mathbf{f}_b^M$ indicates the representation of a PAN image in the MSP feature space. To keep the similarities between HRRS images in different feature spaces, the following equation is formulated:

$$\begin{aligned}\mathcal{L}^T &= \mathcal{L}_c^T + \mathcal{L}_{tri}^T \\ \mathcal{L}_c^T &= \mathrm{CE}(\mathbf{f}_a^P) + \mathrm{CE}(\mathbf{f}_b^M) \\ \mathcal{L}_{tri}^T &= \mathrm{Triplet}\left(\tilde{\mathbf{f}}_a^M, \left(\mathbf{f}_a^P\right)^{pos}, \left(\mathbf{f}_a^P\right)^{neg}, \alpha\right) \\ &+ \mathrm{Triplet}\left(\mathbf{f}_a^P, \left(\tilde{\mathbf{f}}_a^M\right)^{pos}, \left(\tilde{\mathbf{f}}_a^M\right)^{neg}, \alpha\right) \\ &+ \mathrm{Triplet}\left(\tilde{\mathbf{f}}_b^P, \left(\mathbf{f}_b^M\right)^{pos}, \left(\mathbf{f}_b^M\right)^{neg}, \alpha\right) \\ &+ \mathrm{Triplet}\left(\mathbf{f}_b^M, \left(\tilde{\mathbf{f}}_b^P\right)^{pos}, \left(\tilde{\mathbf{f}}_b^P\right)^{neg}, \alpha\right).\end{aligned} \tag{6}$$

In the dual-modality feature learning module, we fuse each source's specific and common features. To explain the fusion process as simply and clearly as possible, we take MSP images as examples. For an MSP image a_i, up to now, we have three representations, they are: the refined specific feature $\tilde{\mathbf{f}}_a^M$, the common feature $\mathbf{f}_a^C$, and the mapped specific feature $\mathbf{f}_a^P$. To describe a_i comprehensively, we fuse $\mathbf{f}_a^C$ with $\tilde{\mathbf{f}}_a^M$ and $\mathbf{f}_a^P$, respectively. In particular, we define

$$\begin{aligned}\mathbf{f}^M &= \omega_{11}\mathbf{f}_a^C + \omega_{12}\tilde{\mathbf{f}}_a^M, \\ \mathbf{f}^{M(P)} &= \omega_{21}\mathbf{f}_a^C + \omega_{22}\mathbf{f}_a^P,\end{aligned} \tag{7}$$

where $\mathbf{f}^M$ denotes the feature of the MSP image a_i in the MSP source, $\mathbf{f}^{M(P)}$ indicates the feature of the MSP image a_i in the PAN source, and ω_{11}, ω_{12}, ω_{21}, and ω_{22} are the weights of different representations that can be learned by the simple fully connected layer. Similarly, for a PAN image b_i, we can obtain its final features $\mathbf{f}^P$ and $\mathbf{f}^{P(M)}$ according to the following fusion scheme:

$$\begin{aligned}\mathbf{f}^P &= \omega_{31}\mathbf{f}_b^C + \omega_{32}\tilde{\mathbf{f}}_b^P, \\ \mathbf{f}^{P(M)} &= \omega_{41}\mathbf{f}_b^C + \omega_{42}\mathbf{f}_b^M.\end{aligned} \tag{8}$$

To further ensure the effectiveness of final features, we use triplet loss to regulate them, i.e.,

$$\begin{aligned}\mathcal{L}^F &= \mathrm{Triplet}\left(\mathbf{f}^M, \left(\mathbf{f}^M\right)^{pos}, \left(\mathbf{f}^M\right)^{neg}, \alpha\right) \\ &+ \mathrm{Triplet}\left(\mathbf{f}^{M(P)}, \left(\mathbf{f}^{M(P)}\right)^{pos}, \left(\mathbf{f}^{M(P)}\right)^{neg}, \alpha\right) \\ &+ \mathrm{Triplet}\left(\mathbf{f}^P, \left(\mathbf{f}^P\right)^{pos}, \left(\mathbf{f}^P\right)^{neg}, \alpha\right) \\ &+ \mathrm{Triplet}\left(\mathbf{f}^{P(M)}, \left(\mathbf{f}^{P(M)}\right)^{pos}, \left(\mathbf{f}^{P(M)}\right)^{neg}, \alpha\right).\end{aligned} \tag{9}$$

3.5. Overall Training and Inference Process

In sum, the overall loss function for training our model is defined as:

$$\mathcal{L}_{overall} = \lambda_1\mathcal{L}^{Sp} + \lambda_2\mathcal{L}^{Cs} + \lambda_3\mathcal{L}^{KL} + \lambda_4\mathcal{L}^T + \lambda_5\mathcal{L}^F. \tag{10}$$

where λ_1, λ_2, λ_3, λ_4, and λ_5 are the hyper-parameters for controlling the contributions of different terms. The training process is summarized in Algorithm 1.

When the DMCL model is trained, we can use it to map the MSP images $\{a_1, a_2, \cdots, a_N\}$ within the archive into $\left\{\mathbf{f}_1^{M(P)}, \mathbf{f}_2^{M(P)}, \cdots, \mathbf{f}_N^{M(P)}\right\}$, and transform PAN images $\{b_1, b_2, \cdots, b_N\}$ within the dataset into $\left\{\mathbf{f}_1^{P(M)}, \mathbf{f}_2^{P(M)}, \cdots, \mathbf{f}_N^{P(M)}\right\}$, respectively. Then, CS-CBRSIR can be completed by measuring the distances between query and target images in the feature space. Particularly, suppose there is an MSP query q^M. We can calculate the distances between its feature $\mathbf{f}_q^M$ and the PAN images' features $\left\{\mathbf{f}_1^{P(M)}, \mathbf{f}_2^{P(M)}, \cdots, \mathbf{f}_N^{P(M)}\right\}$ to search the similar samples from $\{b_1, b_2, \cdots, b_N\}$. Similarly, assume that there is a PAN query q^P. The similarities between its feature $\mathbf{f}_q^P$ and the MSP images' features $\left\{\mathbf{f}_1^{M(P)}, \mathbf{f}_2^{M(P)}, \cdots, \mathbf{f}_N^{M(P)}\right\}$ can be measured to search the similar samples from $\{a_1, a_2, \cdots, a_N\}$. Note that $\mathbf{f}_q^M$ and $\mathbf{f}_q^P$ can be obtained by the trained DCML network directly.

Algorithm 1 Training Process of DMCL.

Input: Dual-source training dataset $D_{train} = \{(P_i, M_i, L_i | i = 1, 2, \cdots, V)\}$, the mini-batch size, the maximum iterations T, and the hyper-parameters λ_1, λ_2, λ_3, λ_4, and λ_5.
Output: The trained DMCL model.
1: Initialize the parameters of our DMCL network.
2: **for** $i = 1, 2, \cdots, T$ **do**
3: Select the triplet datasets from training set for MSP and PAN sources randomly.
4: Obtain the specific features ($\tilde{\mathbf{F}}_a^M$ and $\tilde{\mathbf{F}}_b^P$), common features ($\mathbf{F}_a^C$ and $\mathbf{F}_b^C$), cross-source features ($\mathbf{F}_a^P$, $\mathbf{F}_b^M$), and fusion features ($\mathbf{F}^M$, $\mathbf{F}^{M(P)}$, $\mathbf{F}^P$, and $\mathbf{F}^{P(M)}$) by inputting the triplet datasets into DMCL.
5: Compute the loss value by Equation (10).
6: Update the parameters of the DMCL network by the back propagation algorithm.
7: **end for**

4. Experiments and Analysis

4.1. Experiment Setup

In this paper, we select the DSRSID [51] to verify the proposed method. It comprises 80,000 MSP and PAN image pairs collected from GF-1 optical satellites. The sizes of MSP and PAN are 64×64 and 256×256, and their spatial resolutions are 2 m and 8 m, respectively. MSP images have four spectral channels, while PAN images have single spectral channels. Each image pair covers the same ground region. Those image pairs are equally grouped into eight sematic classes, including "Aquafarm", "Cloud", "Forest", "High Building", "Low Building", "Farm Land", "River", and "Water". The samples of the DSRSID are shown in Figure 3. In the following experiments, we select 75,000 image pairs to train the proposed model, and the remaining 5000 pairs are regarded as queries.

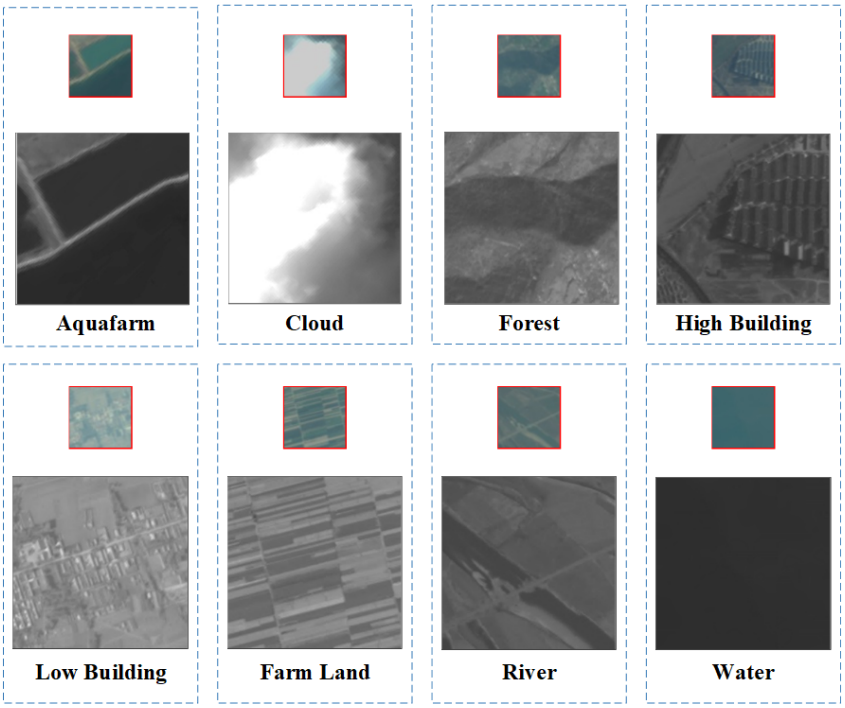

Figure 3. Examples of the DSRSID. The upper example in each block is an MSP image, and the lower sample is a PAN image.

All experiments are conducted on a high-performance computer with GeForce GTX TITAN GPU and 10 G memory. The backbone of our model is initialized with the pre-trained weights provided by [56], and other parts are initialized randomly. The Adam optimizer is adopted to update the parameters. The learning rate is set to be 0.0001 initially, and it is decayed by 0.1 every two epochs. The batch size and training epoch are equal to 48 and 6. The input data are constructed in the triplet format, including anchors and their positive and negative samples. For PAN images, their channels are copied into four to create four-channel images. In addition, the dimensions of immediate and final features are 200 here. As said in Section 3, there are several free parameters in our model, including the margin parameter α within the triplet loss (see Equation (1)), and the hyper-parameters $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5\}$ within the overall loss function (see Equation (10)). In the following experiments, their values equal 0.5 and $\{0.5, 1.0, 0.7, 0.9, 1.0\}$ unless otherwise specified. Their influence on our model will be discussed in Section 4.4.

The retrieval precision–recall curve, precision at k ($P@k$), and the mean average precision (MAP) are selected to evaluate the performance of our model numerically. Precision and recall mean the proportion of correct retrieval results in the returned samples and the truth samples corresponding to the query, respectively. $P@k$ implies the percentage of correct retrieval results when the number of returned samples equals k. The definition of MAP is:

$$\mathrm{MAP} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{N(q_i)}\sum_{j=1}^{N} precision(j)\delta(j), \tag{11}$$

where $q_i \in Q$ is the query image, $|Q|$ is the volume of the query set, $N(q_i)$ indicates the size of ground truth in the target set, N denotes the number of returned samples, $precision(j)$ implies the retrieval precision of the top j retrieved results, and $\delta(j)$ is a Boolean flag parameter that equals 1 when the j-th retrieved sample is correct.

4.2. Performance of DMCL

4.2.1. Reasonableness of Backbone

As said in Section 3.2, the backbone of the specific feature extractor is ResNet50. Before comparing our DCML with other CS-CBRSIR models, we first study if this selection is reasonable or not. To this end, we construct different DMCLs to accomplish the CS-CBRSIR tasks, whose backbones are Alexnet [45], VGG16Net [63], and DenseNet [64]. We record them as DMCL-A, DMCL-V, and DMCL-D, respectively. The MAP values of them and the original DMCL are summarized in Table 1. Additionally, the FLOPs and parameter volumes of different backbones are exhibited for reference. We can find that DMCL achieves the best performance. In addition, the FLOPs and parameter volume of ResNet50 are good compared with the other three backbones.

Table 1. Retrieval results of different DCMLs with different backbones. MUL → PAN means the query is MSP images, and the target samples are PAN images. PAN → MUL indicates the query is PAN images, and the target samples are MSP images.

	DMCL-A	**DMCL-V**	**DMCL-D**	**DMCL**
MUL → PAN MAP (%)	96.57	97.76	98.94	**99.08**
PAN → MUL MAP (%)	97.08	98.06	98.97	**99.32**
Backbone FLOPs (GB)	0.71	15.62	4.28	**3.53**
Backbone Parameters (MB)	61.1	138.36	**20.01**	25.56

4.2.2. Compared with Diverse Methods

In this section, we select twelve methods to verify the proposed DMCL, they are: a deep visual–audio network (DVAN) [48], three models proposed in [65] (one-stream, two-steam, and zero-padding networks), two two-stream networks with and without hierarchical cross-modality metric learning (TONE and TONE + HCML) [66], two dual-path networks introduced in [67] (BCTR and BDTR), a source-invariant deep hashing convolutional neural network (SIDHCNN) [51], a discriminative distillation network (Distillation-ResNet50) [52], an ensemble learning and knowledge distillation network (ELKDN) [68], and a cross-modality shared-specific feature transfer (cm-SSFT) network [18]. Note that the hash code length in SIDHCNN is 32.

The results of different models are exhibited in Table 2, in which the MUL → PAN means that users use MSP images as queries to find similar PAN images and PAN → MUL indicates the opposing case. It is easy for us to find that our DMCL outperforms other methods in all cases. Particularly, for the MUL → PAN task, the MAP enhancements obtained by DMCL are 1.17% (over ELKDN), 1.14% (over Distillation-Res50), 1.68% (over SIDHCNNs), 2.99% (over BDTR), 5.51% (over cm-SSFT), 5.71% (over BCTR), 16.02% (over TONE + HCML), 18.31% (over zero-padding), 19.62% (over one-stream), 20.61% (over TONE), 22.45% (over two-stream), and 25.48% (over DVAN). For the PAN → MUL task, the MAP improvements achieved by DMCL are 0.87% (over ELKDN), 1.23% (over Distillation-Res50), 2.27% (over SIDHCNNs), 3.04% (over BDTR), 4.49% (over cm-SSFT), 7.08% (over BCTR), 16.05% (over TONE + HCML), 19.77% (over zero-padding), 21.58% (over one-stream), 21.81% (over TONE), 23.48% (over two-stream), and 26.60% (over DVAN). Such results are attributed to the following points: (i) the useful specific and common features can be fully extracted by the feature extractor and the common feature learning block, (ii) the obtained features can be fused well by the adaptive dual-modality fusion block, and (iii) both the specific and common features from different RS images are considered simultaneously during the retrieval process. In addition, we show the precision–recall curves of different methods in Figure 4. From observing these curves, we can further discover the superiority of DMCL.

Table 2. MAP values (%) of different models for the DSRSID. MUL → PAN means the query is MSP images, and the target samples are PAN images. PAN → MUL indicates the query is PAN images, and the target samples are MSP images.

Methods	MUL → PAN	PAN → MUL
DMCL (Ours)	**99.08**	**99.32**
ELKDN [68]	97.91	98.45
Distillation-Res50 [52]	97.94	98.09
SIDHCNNs [51]	97.40	97.05
BDTR [67]	96.09	96.28
cm-SSFT [18]	93.57	94.83
BCTR [67]	93.37	92.24
TONE + HCML [66]	83.06	83.27
Zero-padding [65]	80.77	79.55
One-stream [65]	79.46	77.74
TONE [66]	78.47	77.51
Two-stream [65]	76.63	75.84
DVAN [48]	73.60	72.72

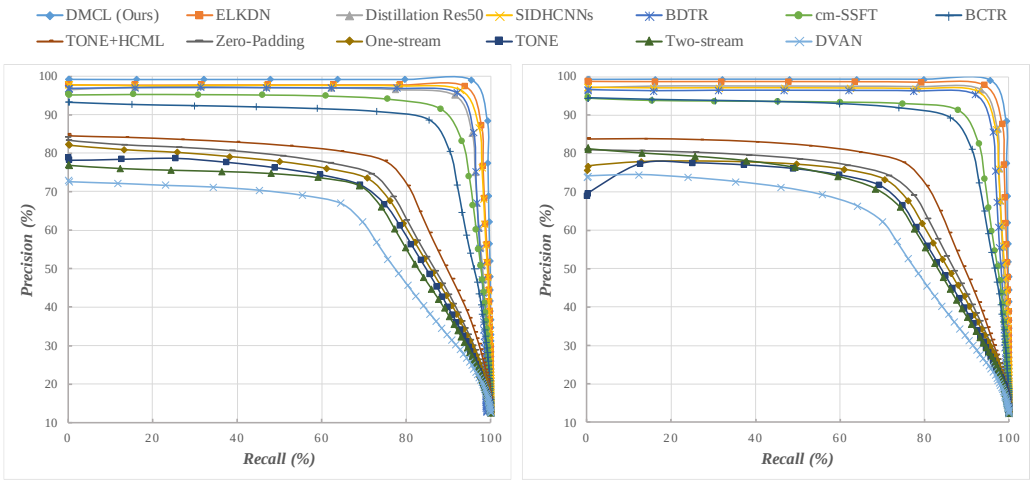

Figure 4. Comparisons of precision–recall curve of different methods.(**a**) Comparisons of different methods in MUL → PAN task. (**b**) Comparisons of different methods in PAN → MUL task.

To further study DMCL, we measure its performance on different semantic classes in Table 3. Other methods' results are also displayed for reference. From the observation of these results, we can find that the performance of DMCL is the best among all of the methods. Taking the MUL → PAN task as an example, the largest increasing ranges between DCML and other methods are concentrated upon "High Building" (over ELKDN, Distillation-Res50, TONE + HCML, TONE, and two-stream), "Aquafarm" (over cm-SSFT, BCTR, one-stream, and DVAN), "Forest" (over SIDHCNNs and zero-padding), and "Cloud" (over BDTR). An encouraging observation is that the MAP values of DMCL in "Cloud" and "Water" are as high as 100%. Such good results are due to the following two factors. One is that the HRRS images from "Cloud" and "Water" have relatively simple contents. Thus, recognizing them correctly is not a difficult issue. The other is that the features of different HRRS images captured by DMCL are representative and informative enough. Therefore, their similarity relationships can be measured exactly. However, we should note that the performance gap between DMCL and ELKDN is not as large as expected. The main reason behind this is that their key ideas are similar, i.e., finding the proper common space for HRRS images from different sources. Although our DMCL adds an adaptive dual-modality fusion block to supplement the specific information to common features, the related operations (e.g., modality transformation) do not fully consider the characteristics of HRRS. Therefore, DMCL outperforms ELKDN slightly. How to enhance the dual-modality fusion block is one of our future works.

Apart from the numerical results, the visual retrieval results are also exhibited in Figures 5 and 6. The queries are randomly selected from the DSRSID and shown in the top block. Then, the corresponding retrieval samples and their ranks are exhibited in the bottom block. In addition, the incorrect retrieval results are tagged in red for clarity. We can find that most of the top-ranked retrieved results are correct, which illustrates that our model is useful for CS-CBRSIR tasks.

The positive results discussed in this section confirm the first contribution summarized in Section 1, i.e., DMCL is good at dealing with CS-CBRSIR.

Table 3. MAP values (%) of different models measured in the DSRSID. MUL → PAN means the query is MSP images, and the target samples are PAN images. PAN → MUL indicates the query is PAN images, and the target samples are MSP images.

Task	Methods	Aquafarm	Cloud	Forest	High Building	Low Building	Farm Land	River	Water
MUL → PAN	**DMCL (Ours)**	**99.17**	**100**	**98.01**	**99.41**	**98.69**	**98.07**	**99.28**	**100**
	ELKDN [68]	97.24	100	97.05	96.77	98.04	96.16	98.37	99.66
	Distillation-Res50 [52]	97.34	99.93	96.99	96.81	97.99	96.10	98.61	99.80
	SIDHCNNs [51]	96.73	99.97	94.71	96.63	96.22	97.16	98.03	99.74
	BDTR [67]	93.04	92.29	95.44	95.91	96.83	97.34	97.85	100
	cm-SSFT [18]	88.63	99.82	93.21	90.06	92.68	88.79	95.95	99.40
	BCTR [67]	89.10	99.17	88.95	89.43	88.68	97.77	94.56	99.32
	TONE + HCML [66]	64.13	98.26	82.61	62.61	81.88	88.01	90.53	96.45
	Zero-padding [65]	70.34	99.72	59.27	69.94	78.34	87.60	81.29	99.67
	One-stream [65]	61.70	99.19	73.97	62.42	71.09	85.71	82.86	98.74
	TONE [66]	64.98	79.28	71.02	58.8	82.48	84.80	86.46	86.46
	Two-stream [65]	61.38	96.99	71.50	53.83	74.56	73.76	85.43	95.63
	DVAN [48]	58.31	94.55	63.84	59.44	66.41	71.20	80.78	94.47
PAN → MUL	**DMCL (Ours)**	**99.64**	**100**	**97.76**	**99.26**	**99.25**	**99.22**	**99.45**	**99.97**
	ELKDN [68]	98.45	100	96.41	97.97	98.01	97.86	99.05	99.87
	Distillation-Res50 [52]	98.12	99.61	95.84	97.60	97.63	97.68	98.78	99.50
	SIDHCNNs [51]	95.61	99.98	94.33	94.33	96.8	96.71	97.95	99.90
	BDTR [67]	97.15	91.35	95.29	97.56	93.62	96.92	98.62	99.74
	cm-SSFT [18]	92.95	99.69	87.73	93.89	91.18	96.94	96.69	99.55
	BCTR [67]	86.44	100	84.10	86.35	94.27	92.05	94.80	99.95
	TONE + HCML [66]	77.37	97.61	62.70	71.19	81.05	87.37	90.11	98.77
	Zero-padding [65]	68.80	99.48	65.14	70.96	71.36	78.49	85.26	96.92
	One-stream [65]	63.07	97.71	72.11	64.63	63.84	76.51	86.37	97.68
	TONE [66]	56.63	99.21	71.51	69.07	74.22	86.61	77.12	85.75
	Two-stream [65]	62.19	95.78	68.27	54.69	71.23	75.40	84.36	94.79
	DVAN [48]	59.63	93.46	61.24	54.69	65.79	73.20	79.95	93.83

4.3. Ablation Study

As discussed in Section 3, the proposed DCML consists of a specific feature extractor, a common feature learning block, and an adaptive dual-modality fusion block. They aim to extract specific, common, and fused features from RS images corresponding to different sources. To study their contributions to DCML, we first construct three networks in this section, i.e.,

- Net1: specific feature extractor,
- Net2: specific feature extractor + common feature learning block,
- Net3: specific feature extractor + common feature learning block + adaptive dual-modality fusion block.

Then, three networks are applied on DSRSID to measure the performance summarized in Table 4. Note that the experimental settings are the same as the contents mentioned in Section 4.1. From observing the results, we find that the common feature learning and adaptive dual-modality fusion blocks make positive contributions based on the specific feature extractor. For instance, in the MUL → PAN task, when there is only the specific feature extractor, the *P*@1, *P*@3000, *P*@8000, and MAP values of retrieval results are merely 96.98%, 96.92%, 96.86%, and 96.91%. This demonstrates that the distribution differences of specific features are too large to calculate the similarity relations. Once the common

feature learning block is added, the behavior of Net2 becomes distinctly stronger. Its assessment criteria values are as high as 98.61% (*P*@1), 98.63% (*P*@3000), 98.59% (*P*@8000), and 98.63% (MAP), which indicates the effectiveness of the common feature learning block. In other words, since the modality discrepancy has been narrowed down, the common features can be used to measure the similarities between different RS images (i.e., the second contribution of DMCL). Even though the performance of Net2 is adequate, the results of Net3 are more satisfactory than those of Net2, which implies that integrating specific and common features is beneficial to CS-CBRSIR (i.e., the third contribution of DMCL).

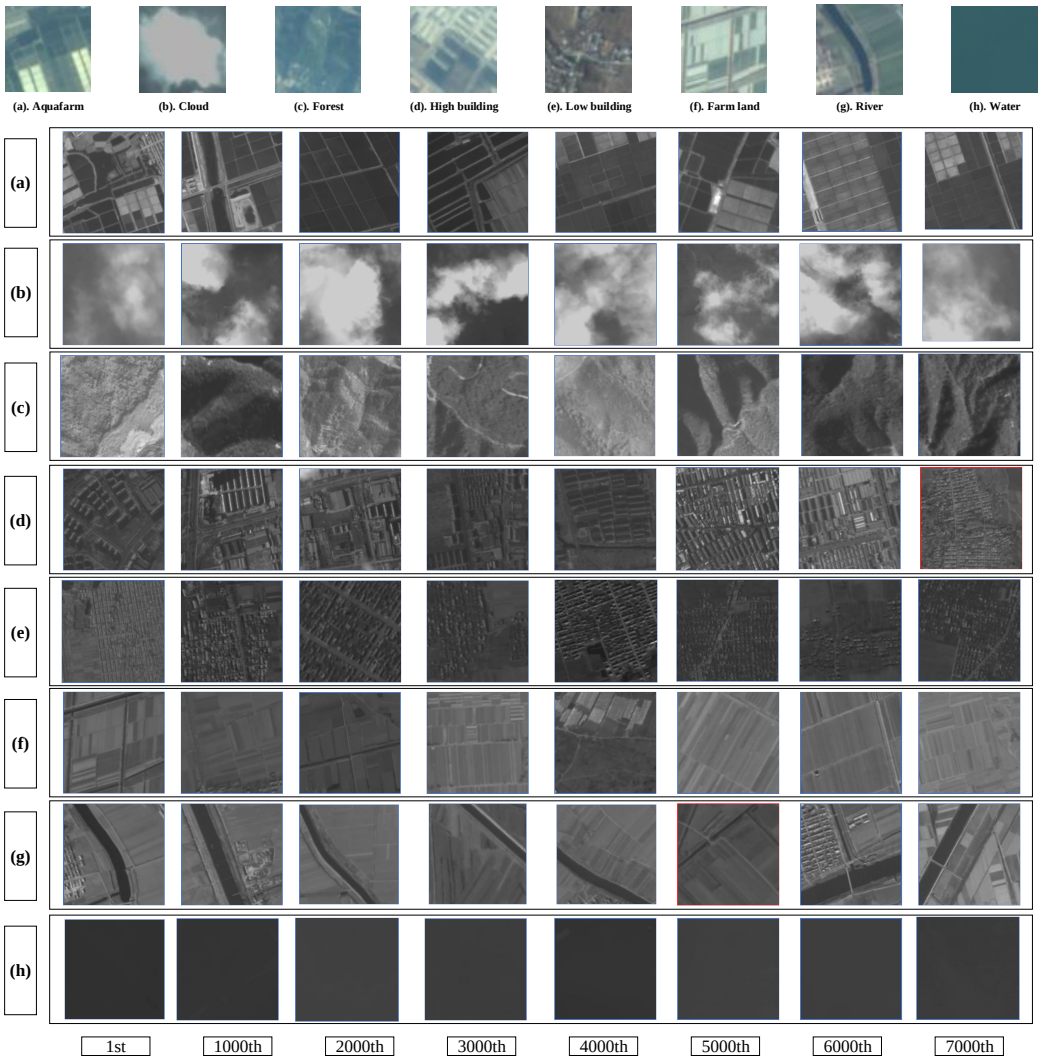

Figure 5. Visual retrieval results on cross-source MUL → PAN. Inquiry images in the first row are from a panchromatic source and they are picked from each class respectively. (**a**–**h**) are retrieved images for the corresponding inquiry images in the first row. The red rectangles denote the false retrieved images that are irrelevant to inquiry images.

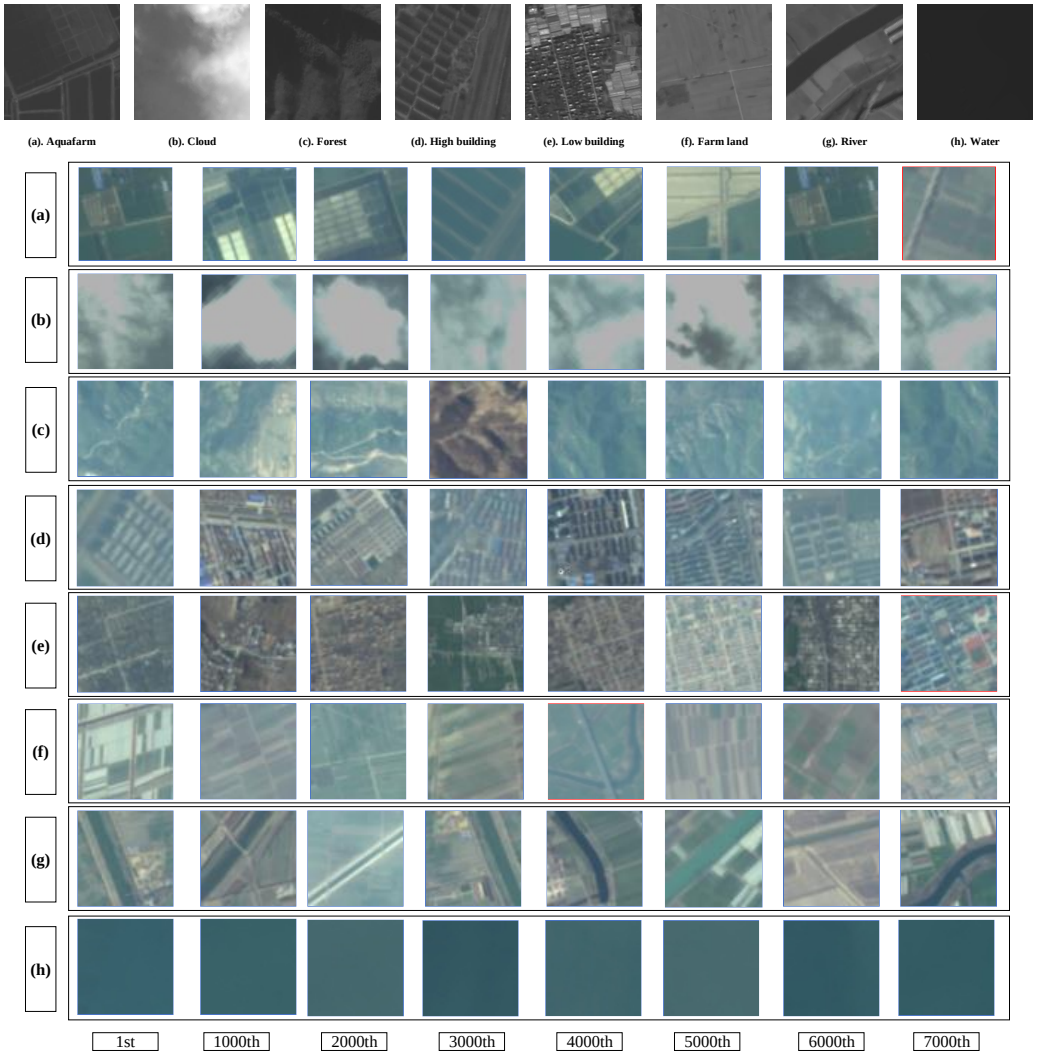

Figure 6. Visual retrieval results on cross-source PAN → MUL. Inquiry images in the first row are from a panchromatic source and they are picked from each class respectively. (**a**–**h**) are retrieved images for the corresponding inquiry images in the first row. The red rectangles denote the false retrieved images that are irrelevant to inquiry images.

Table 4. Performance of different blocks.

Task	Networks	*P*@1	*P*@3000	*P*@8000	MAP
MUL → PAN	Net1	96.98	96.92	96.86	96.91
	Net2	98.61	98.63	98.59	98.63
	Net3	99.08	99.09	99.02	99.08
PAN → MUL	Net1	96.48	96.44	96.35	96.41
	Net2	98.76	99.25	99.13	98.75
	Net3	99.34	99.38	99.19	99.32

4.4. Sensitive Study

In this section, we study the sensitivity of our DMCL from two aspects. One is the impact of feature dimensions, and the other is the influence of free parameters.

As discussed in Section 4.1, we set the dimensions of immediate and final features to be 200. To study the influence of different feature dimensions, we change their length

from 100 to 400 with an interval of 100. Based on those features, we can construct different DCMLs based on those features. Their MAP curves are exhibited in Figure 7a. From the observation of them, we can find that there is a distinct performance gap between DCMLs with 100- and 200-dimensional features. When the dimension is higher than 200, the performance of DCMLs increases slightly. Taking the performance and time costs into account simultaneously, we set the dimensions of features to 200 in this study.

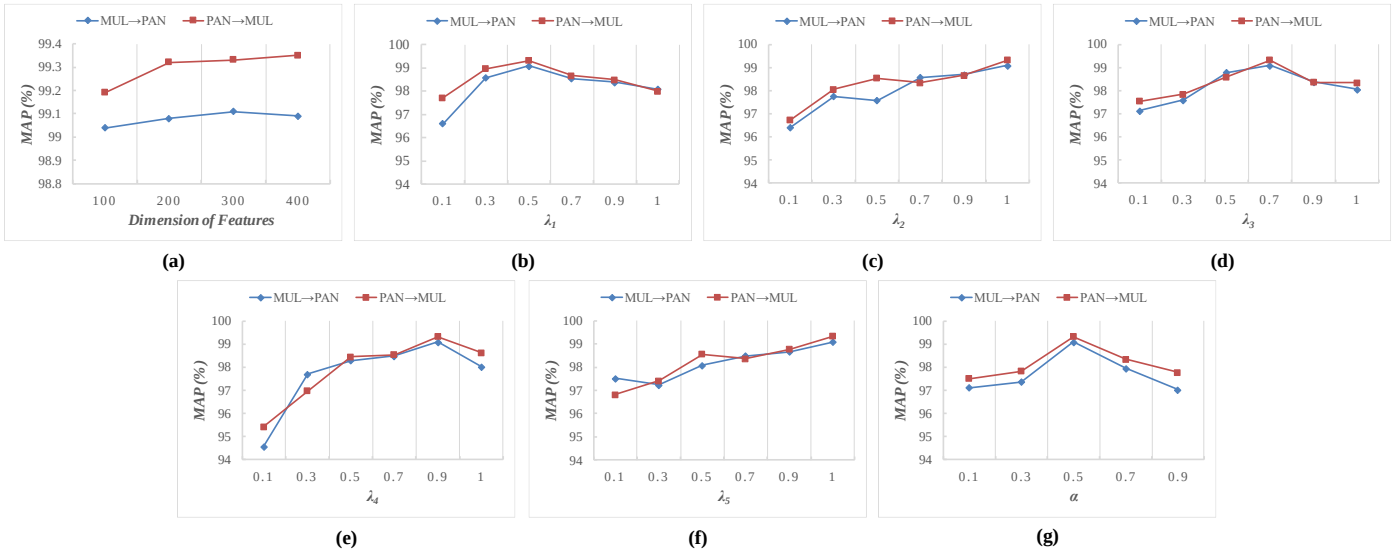

Figure 7. Influence of feature dimensions and different free parameters. (**a**) Feature dimensions. (**b**) λ_1. (**c**) λ_2. (**d**) λ_3. (**e**) λ_4. (**f**) λ_5. (**g**) α.

There are six free parameters in the proposed method, including the hyper-parameters $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5\}$ and the margin parameter α. To study their influence on DCML, we vary their values within certain ranges. For clarity, only one parameter is changed at once and the others are equal to the values mentioned in Section 4.1. For $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5\}$, we vary their values from 0.1 to 1 with an interval of 0.2. The performance changes in different DMCLs based on various hyper-parameters are exhibited in Figure 7b–f. It is easy to find that the performance of DMCLs is improved with λ_1, λ_3, and λ_4 increasing. The peak values appear at $\lambda_1 = 0.5$, $\lambda_3 = 0.7$, and $\lambda_4 = 0.9$. Then, their behavior becomes weaker. Meanwhile, the performance of DMCLs is enhanced with λ_2 and λ_5 increasing, and the peak values appear at $\lambda_2 = 1.0$ and $\lambda_5 = 1.0$. For α, its values are changed from 0.1 to 0.9 with an interval of 0.2, and its impacts on DMCL are shown in Figure 7g. We can find that the optimal performance occurs at $\alpha = 0.5$. When $\alpha > 0.5$ or $\alpha < 0.5$, the performance of our model declines.

Note that the above sensitive study is general purpose, and the curves shown in Figure 7 are measured in the DSRSID. If readers use some new datasets, the optimal parameters and their behavior variation tendency would change.

5. Conclusions

This paper proposes a new model for CS-CSRSIR tasks named DMCL. First, we select ResNet50 as the backbone to extract the specific features from HRRS images. Then, the common mutual learning module is developed to map the specific features into a common space so that the discrepancy of diverse modalities can be reduced. Next, to further improve the representation for different HRRS images, we introduce the modality transform scheme and dual-space feature fusion module. They supplement the specific information to common features and fuse the knowledge from various sources. The positive experimental results demonstrate that our DMCL is helpful in CS-CSRSIR tasks. Apart from CS-CBRSIR, our DMCL has the potential to complete other computer vision tasks. The critical points of DMCL are the modality transformation and feature consistency learning.

Those two techniques are also the core technologies in various cross-modal applications, such as audio–visual generation, vision-and-language pre-training, etc. Therefore, the framework of DMCL can be transferred to various applications as long as researchers make proper adjustments according to different demands.

Although the proposed method achieves encouraging results, there is still room for improvement. First, rather than selecting a general-purpose model (ResNet50), developing an HRRS-oriented network as our backbone to learn specific features from HRRS images will benefit our tasks. Second, some particular constraints could be proposed and added to the modality transition matrix W_t considering the characteristics of HRRS images such that the proposed method will fit the RS scenario better. How to deal with the limitation mentioned above is our future work.

Author Contributions: Conceptualization, X.T. and D.S.; methodology, D.S. and J.M.; software, D.S.; writing—original draft preparation, X.T. and D.S.; writing—review and editing, X.Z. and L.J.; funding acquisition, J.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded in part by the National Natural Science Foundation of China (No. 62171332), Key Research and Development Program of Shaanxi (Nos. 2021GY-035 and 2019ZDLGY03-08), Fund of National Key Laboratory of Science and Technology on Remote Sensing Information and imagery Analysis, Beijing Research Institute of Uranium Geology (No. 6142A010301), China Postdoctoral Science Foundation Funded Project (No. 2017M620441), and Key Laboratory Program (No. HLJGXQ20210701008).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Li, Y.; Ma, J.; Zhang, Y. Image retrieval from remote sensing big data: A survey. *Inf. Fusion* **2021**, *67*, 94–115. [CrossRef]
2. Tang, X.; Zhang, H.; Mou, L.; Liu, F.; Zhang, X.; Zhu, X.X.; Jiao, L. An Unsupervised Remote Sensing Change Detection Method Based on Multiscale Graph Convolutional Network and Metric Learning. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5609715. [CrossRef]
3. Yang, Y.; Tang, X.; Cheung, Y.M.; Zhang, X.; Liu, F.; Ma, J.; Jiao, L. AR 2 Det: An Accurate and Real-Time Rotational One-Stage Ship Detector in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5605414.
4. Tang, X.; Meng, F.; Zhang, X.; Cheung, Y.M.; Ma, J.; Liu, F.; Jiao, L. Hyperspectral image classification based on 3-D octave convolution with spatial–spectral attention network. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 2430–2447. [CrossRef]
5. González-Briones, A.; Prieto, J.; De La Prieta, F.; Demazeau, Y.; Corchado, J.M. Virtual agent organizations for user behaviour pattern extraction in energy optimization processes: A new perspective. *Neurocomputing* **2021**, *452*, 374–385. [CrossRef]
6. Decuyper, M.; Chávez, R.O.; Lohbeck, M.; Lastra, J.A.; Tsendbazar, N.; Hackländer, J.; Herold, M.; Vågen, T.G. Continuous monitoring of forest change dynamics with satellite time series. *Remote Sens. Environ.* **2022**, *269*, 112829. [CrossRef]
7. Xu, X.; Zhang, L.; Trovati, M.; Palmieri, F.; Asimakopoulou, E.; Johnny, O.; Bessis, N. PERMS: An efficient rescue route planning system in disasters. *Appl. Soft Comput.* **2021**, *111*, 107667. [CrossRef]
8. Tong, X.Y.; Xia, G.S.; Hu, F.; Zhong, Y.; Datcu, M.; Zhang, L. Exploiting deep features for remote sensing image retrieval: A systematic investigation. *IEEE Trans. Big Data* **2019**, *6*, 507–521. [CrossRef]
9. Jiao, L.; Tang, X.; Hou, B.; Wang, S. SAR images retrieval based on semantic classification and region-based similarity measure for earth observation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 3876–3891. [CrossRef]
10. Tang, X.; Jiao, L.; Emery, W.J. SAR image content retrieval based on fuzzy similarity and relevance feedback. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 1824–1842. [CrossRef]
11. Tang, X.; Liu, C.; Ma, J.; Zhang, X.; Jiao, L. Large-Scale Remote Sensing Image Retrieval Based on Semi-Supervised Adversarial Hashing. *Remote Sens.* **2019**, *11*, 2055. [CrossRef]
12. Liu, C.; Ma, J.; Tang, X.; Liu, F.; Zhang, X.; Jiao, L. Deep hash learning for remote sensing image retrieval. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 3420–3443. [CrossRef]
13. Tang, X.; Jiao, L. Fusion similarity-based reranking for SAR image retrieval. *IEEE Geosci. Remote Sens. Lett.* **2016**, *14*, 242–246. [CrossRef]
14. Tang, X.; Zhang, X.; Liu, F.; Jiao, L. Unsupervised deep feature learning for remote sensing image retrieval. *Remote Sens.* **2018**, *10*, 1243. [CrossRef]

15. Tang, X.; Yang, Y.; Ma, J.; Cheung, Y.m.; Liu, C.; Liu, F.; Zhang, X.; Jiao, L. Meta-hashing for Remote Sensing Image Retrieval. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5615419. [CrossRef]
16. Tang, X.; Jiao, L.; Emery, W.J.; Liu, F.; Zhang, D. Two-stage reranking for remote sensing image retrieval. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 5798–5817. [CrossRef]
17. Chen, L.; Maddox, R.K.; Duan, Z.; Xu, C. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7832–7841.
18. Lu, Y.; Wu, Y.; Liu, B.; Zhang, T.; Li, B.; Chu, Q.; Yu, N. Cross-modality person re-identification with shared-specific feature transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 13379–13389.
19. Ye, M.; Lan, X.; Leng, Q.; Shen, J. Cross-modality person re-identification via modality-aware collaborative ensemble learning. *IEEE Trans. Image Process.* **2020**, *29*, 9387–9399. [CrossRef]
20. Yu, E.; Ma, J.; Sun, J.; Chang, X.; Zhang, H.; Hauptmann, A.G. Deep Discrete Cross-Modal Hashing with Multiple Supervision. *Neurocomputing* 2021, *in press*. [CrossRef]
21. Zou, X.; Wu, S.; Bakker, E.M.; Wang, X. Multi-label enhancement based self-supervised deep cross-modal hashing. *Neurocomputing* **2022**, *467*, 138–162. [CrossRef]
22. Kaur, P.; Pannu, H.S.; Malhi, A.K. Comparative analysis on cross-modal information retrieval: A review. *Comput. Sci. Rev.* **2021**, *39*, 100336. [CrossRef]
23. Liu, Y.; Chen, Q.; Albanie, S. Adaptive Cross-Modal Prototypes for Cross-Domain Visual-Language Retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 14954–14964.
24. Huang, X.; Peng, Y.; Yuan, M. MHTN: Modal-adversarial hybrid transfer network for cross-modal retrieval. *IEEE Trans. Cybern.* **2018**, *50*, 1047–1059. [CrossRef] [PubMed]
25. Li, X.; Yang, J.; Ma, J. Recent developments of content-based image retrieval (CBIR). *Neurocomputing* **2021**, *452*, 675–689. [CrossRef]
26. Bretschneider, T.; Cavet, R.; Kao, O. Retrieval of remotely sensed imagery using spectral information content. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Toronto, ON, Canada, 24–28 June 2002; Volume 4, pp. 2253–2255.
27. Datcu, M.; Seidel, K.; Walessa, M. Spatial information retrieval from remote-sensing images. I. Information theoretical perspective. *IEEE Trans. Geosci. Remote Sens.* **1998**, *36*, 1431–1445. [CrossRef]
28. Haralick, R.M.; Shanmugam, K.; Dinstein, I.H. Textural features for image classification. *IEEE Trans. Syst. Man Cybern.* **1973**, *SMC-3*, 610–621. [CrossRef]
29. Mallat, S.G. A theory for multiresolution signal decomposition: The wavelet representation. In *Fundamental Papers in Wavelet Theory*; Princeton University Press: Princeton, NJ, USA, 2009; pp. 494–513.
30. Melissaratos, L.; Micheli-Tzanakou, E. Comments on" Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression. *IEEE Trans. Acoust. Speech Signal Process.* **1990**, *38*, 2005. [CrossRef]
31. Scott, G.J.; Klaric, M.N.; Davis, C.H.; Shyu, C.R. Entropy-balanced bitmap tree for shape-based object retrieval from large-scale satellite imagery databases. *IEEE Trans. Geosci. Remote Sens.* **2010**, *49*, 1603–1616. [CrossRef]
32. Ma, A.; Sethi, I.K. Local shape association based retrieval of infrared satellite images. In Proceedings of the Seventh IEEE International Symposium on Multimedia (ISM'05), Irvine, CA, USA, 12–14 December 2005; pp. 551–557.
33. Barros, J.E.; French, J.C.; Martin, W.N.; Kelly, P.M. System for indexing multispectral satellite images for efficient content-based retrieval. In Proceedings of the Storage and Retrieval for Image and Video Databases III. International Society for Optics and Photonics, San Diego/La Jolla, CA, USA, 5–10 February 1995; Volume 2420, pp. 228–237.
34. Shao, Z.; Zhou, W.; Zhang, L.; Hou, J. Improved color texture descriptors for remote sensing image retrieval. *J. Appl. Remote Sens.* **2014**, *8*, 083584. [CrossRef]
35. Yang, J.; Liu, J.; Dai, Q. An improved Bag-of-Words framework for remote sensing image retrieval in large-scale image databases. *Int. J. Digit. Earth* **2015**, *8*, 273–292. [CrossRef]
36. Perronnin, F.; Dance, C. Fisher kernels on visual vocabularies for image categorization. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 18–23 June 2007; pp. 1–8.
37. Jégou, H.; Douze, M.; Schmid, C.; Pérez, P. Aggregating local descriptors into a compact image representation. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 3304–3311.
38. Yang, Y.; Newsam, S. Geographic image retrieval using local invariant features. *IEEE Trans. Geosci. Remote Sens.* **2012**, *51*, 818–832. [CrossRef]
39. Aptoula, E. Bag of morphological words for content-based geographical retrieval. In Proceedings of the 2014 12th International Workshop on Content-Based Multimedia Indexing (CBMI), Klagenfurt, Austria, 18–20 June 2014; pp. 1–5.
40. Bosilj, P.; Aptoula, E.; Lefèvre, S.; Kijak, E. Retrieval of remote sensing images with pattern spectra descriptors. *ISPRS Int. J. Geo-Inf.* **2016**, *5*, 228. [CrossRef]
41. Sun, Y.; Ye, Y.; Li, X.; Feng, S.; Zhang, B.; Kang, J.; Dai, K. Unsupervised deep hashing through learning soft pseudo label for remote sensing image retrieval. *Knowl.-Based Syst.* **2021**, *239*, 107807. [CrossRef]
42. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]

43. Dewi, C.; Chen, R.C.; Yu, H. Weight analysis for various prohibitory sign detection and recognition using deep learning. *Multimed. Tools Appl.* **2020**, *79*, 32897–32915. [CrossRef]
44. Zhou, W.; Newsam, S.; Li, C.; Shao, Z. Learning Low Dimensional Convolutional Neural Networks for High-Resolution Remote Sensing Image Retrieval. *Remote Sens.* **2016**, *9*, 489. [CrossRef]
45. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*. [CrossRef]
46. Makhzani, A.; Shlens, J.; Jaitly, N.; Goodfellow, I.; Frey, B. Adversarial autoencoders. *arXiv* **2015**, arXiv:1511.05644.
47. Yang, M.Y.; Landrieu, L.; Tuia, D.; Toth, C. Muti-modal learning in photogrammetry and remote sensing. *ISPRS J. Photogramm. Remote Sens.* **2021**, *176*, 54. [CrossRef]
48. Mao, G.; Yuan, Y.; Xiaoqiang, L. Deep cross-modal retrieval for remote sensing image and audio. In Proceedings of the 2018 10th IAPR Workshop on Pattern Recognition in Remote Sensing (PRRS), Beijing, China, 19–20 August 2018; pp. 1–7.
49. Lu, X.; Wang, B.; Zheng, X.; Li, X. Exploring models and data for remote sensing image caption generation. *IEEE Trans. Geosci. Remote Sens.* **2017**, *56*, 2183–2195. [CrossRef]
50. Yuan, Z.; Li, X.; Wang, Q. Exploring Multi-Level Attention and Semantic Relationship for Remote Sensing Image Captioning. *IEEE Access* **2019**, *8*, 2608–2620. [CrossRef]
51. Li, Y.; Zhang, Y.; Huang, X.; Ma, J. Learning source-invariant deep hashing convolutional neural networks for cross-source remote sensing image retrieval. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 6521–6536. [CrossRef]
52. Xiong, W.; Xiong, Z.; Cui, Y.; Lv, Y. A Discriminative Distillation Network for Cross-Source Remote Sensing Image Retrieval. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 1234–1247. [CrossRef]
53. Chaudhuri, U.; Banerjee, B.; Bhattacharya, A.; Datcu, M. CMIR-NET: A deep learning based model for cross-modal retrieval in remote sensing. *Pattern Recognit. Lett.* **2020**, *131*, 456–462. [CrossRef]
54. Xiong, W.; Lv, Y.; Zhang, X.; Cui, Y. Learning to Translate for Cross-Source Remote Sensing Image Retrieval. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 4860–4874. [CrossRef]
55. Schroff, F.; Kalenichenko, D.; Philbin, J. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 815–823.
56. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
57. Anwer, R.M.; Khan, F.S.; Van De Weijer, J.; Molinier, M.; Laaksonen, J. Binary patterns encoded convolutional neural networks for texture recognition and remote sensing scene classification. *ISPRS J. Photogramm. Remote Sens.* **2018**, *138*, 74–85. [CrossRef]
58. Zhang, X.; Wang, G.; Zhu, P.; Zhang, T.; Li, C.; Jiao, L. GRS-Det: An anchor-free rotation ship detector based on Gaussian-mask in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 3518–3531. [CrossRef]
59. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
60. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
61. He, X.; Ma, W.Y.; Zhang, H.J. Learning an image manifold for retrieval. In Proceedings of the 12th Annual ACM International Conference on Multimedia, New York, NY, USA, 10–16 October 2004; pp. 17–23.
62. Tian, Y.; Fan, B.; Wu, F. L2-net: Deep learning of discriminative patch descriptor in euclidean space. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 661–669.
63. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
64. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
65. Wu, A.; Zheng, W.S.; Yu, H.X.; Gong, S.; Lai, J. RGB-infrared cross-modality person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5380–5389.
66. Ye, M.; Lan, X.; Li, J.; Yuen, P. Hierarchical discriminative learning for visible thermal person re-identification. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
67. Ye, M.; Wang, Z.; Lan, X.; Yuen, P.C. Visible thermal person re-identification via dual-constrained top-ranking. In Proceedings of the IJCAI, Stockholm, Sweden, 13–19 July 2018; Volume 1, p. 2.
68. Ma, J.; Shi, D.; Tang, X.; Zhang, X.; Han, X.; Jiao, L. Cross-Source Image Retrieval Based on Ensemble Learning and Knowledge Distillation for Remote Sensing Images. In Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, 12–16 July 2021; pp. 2803–2806.