Informative and Representative Triplet Selection for Multilabel Remote Sensing Image Retrieval

Gencer Sumbul, *Graduate Student Member, IEEE*, Mahdyar Ravanbakhsh, *Member, IEEE*, and Begüm Demir, *Senior Member, IEEE*

***Abstract*— Learning the similarity between remote sensing (RS) images forms the foundation for content-based RS image retrieval (CBIR). Recently, deep metric learning approaches that map the semantic similarity of images into an embedding (metric) space have been found very popular in RS. A common approach for learning the metric space relies on the selection of triplets of similar (positive) and dissimilar (negative) images to a reference image called an anchor. Choosing triplets is a difficult task particularly for multilabel RS CBIR, where each training image is annotated by multiple class labels. To address this problem, in this article, we propose a novel triplet sampling method in the framework of deep neural networks (DNNs) defined for multilabel RS CBIR problems. The proposed method selects a small set of the most representative and informative triplets based on two main steps. In the first step, a set of anchors that are diverse to each other in the embedding space is selected from the current minibatch using an iterative algorithm. In the second step, different sets of positive and negative images are chosen for each anchor by evaluating the relevancy, hardness, and diversity of the images among each other based on a novel strategy. Experimental results obtained on two multilabel benchmark archives show that the selection of the most informative and representative triplets in the context of DNNs results in: 1) reducing the computational complexity of the training phase of the DNNs without any significant loss on the performance and 2) an increase in learning speed since informative triplets allow fast convergence. The code of the proposed method is publicly available at https://git.tu-berlin.de/rsim/image-retrieval-from-triplets.**

***Index Terms*— Deep neural networks (DNNs), metric learning, multilabel image retrieval, remote sensing (RS), triplet selection.**

I. Introduction

In recent years, advancements in satellite technology have led to fast-growing archives of remote sensing (RS) images. One of the most emerging applications in RS is the accurate retrieval of RS images from such archives. Thus, the development of content-based image retrieval (CBIR) methods has recently attracted great attention [1]. The performance of any CBIR system relies on its capability to learn discriminative and robust image representations to describe the complex semantic content of RS images.

Conventional CBIR systems exploit handcrafted features to describe the content of images. As an example, Wang and Newsam present a retrieval system employing the well-known scale-invariant feature transform (SIFT) to extract bag-of-visual-words representations of image features [2]. Aptoula introduces the use of bag-of-morphological-words representations for local texture descriptors [3]. In [4], a comparative analysis of local binary patterns (LBPs) that capture local patterns between neighboring pixels is presented. Chaudhuri *et al.* [5] present a method that represents image content by a graph, where the graph nodes describe the image region properties and the edges represent the spatial relationships among the regions. Binary hash codes obtained through kernel-based hashing methods are found effective for describing RS images in [6]. After extracting the image features, the most similar images with respect to a query image can be found by performing the k-nearest neighbor (k-NN) search algorithm. In the case of graph-based image representations, graph comparison methods, such as the inexact graph matching approach, proposed by Chaudhuri *et al.* [7] can be used. The images represented by binary hash codes can be searched and retrieved by using the computationally efficient hamming distance [6].

The abovementioned CBIR systems cannot simultaneously optimize feature learning and image retrieval and, thus, result in a limited capability to represent the high-level semantic content of RS images. This issue leads to insufficient search and retrieval performance [1]. To overcome this problem, CBIR systems based on deep neural networks (DNNs) have been recently presented in RS [8]. As an example, Li *et al.* [9] propose a method that fuses deep features and handcrafted features. This method exploits four convolutional neural networks (CNNs) to extract features at different steps and with different coarse levels. Then, these deep features are fused with traditional image descriptors, such as LBPs and SIFT, to be used in the retrieval process. A convolutional autoencoder is used by Tang *et al.* [10] to obtain deep bag-of-words image descriptors. To this end, a reconstruction loss function that minimizes the error between the input and the extracted descriptors is considered. Imbriaco *et al.* [11] extract local convolutional features and aggregate them into a global descriptor, where the deep features are extracted through a pretrained model without any fine-tuning. Boualleg and Farrah address

Manuscript received May 7, 2021; revised September 23, 2021; accepted October 23, 2021. Date of publication October 29, 2021; date of current version February 14, 2022. This work was supported in part by the European Research Council (ERC) through the ERC-2017-STG BigEarth Project under Grant 759764 and in part by the German Research Foundation as part of the priority program "Volunteered Geographic Information: Interpretation, Visualization and Social Computing" (VGIscience, priority program 1894). *(Corresponding author: Begüm Demir.)*

The authors are with the Faculty of Electrical Engineering and Computer Science, Technische Universität Berlin, 10623 Berlin, Germany (e-mail: gencer.suembuel@tu-berlin.de; ravanbakhsh@tu-berlin.de; demir@tu-berlin.de).

Digital Object Identifier 10.1109/TGRS.2021.3124326

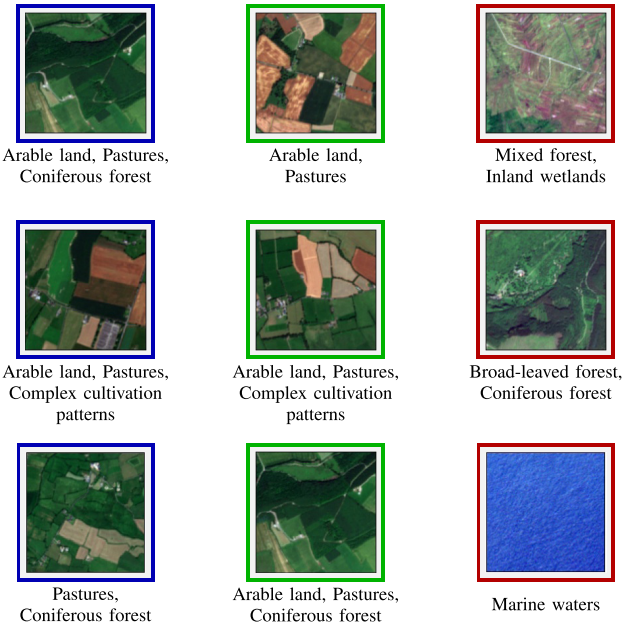

Fig. 1. Example of three triplets consisting of images from BigEarthNet [15]. Each triplet given in different rows consists of an anchor (in blue frame), a positive image (in green frame), and a negative image (in red frame). The associated multilabels are given below the respective images.

the semantic gap between low-level features and high-level perception of semantic similarity in [12]. This is achieved by using a CNN to detect semantic concepts and a relevance feedback strategy to ensure that CBIR results match with a query image. Sabahi *et al.* [13] address the abovementioned semantic gap by employing a recurrent neural network to model the human visual memory.

In recent years, deep metric learning (DML)-based methods that aim at learning a feature space (in which similar images are close to each other) have attracted attention in RS. Current DML models are mostly trained using a triplet loss function made up of three images as: 1) an *anchor image*; 2) a *positive image* that is similar to the anchor; and 3) a *negative image* that is dissimilar to the anchor [14]. An example of triplets constructed from BigEarthNet [15] can be seen in Fig. 1. A difficult task in DML is to construct the set of triplets. A simple strategy is to define triplets from an existing training set of labeled images. Roy *et al.* [16] apply a strategy that: 1) randomly selects an anchor from a minibatch of training images and 2) randomly chooses one positive image that has the same class label as the anchor while selecting one negative image that has a different class label. Similarly, Lai *et al.* [8] select triplets randomly based on the class labels of training images to train an end-to-end model for hashing. For each anchor image, there can be several positive and negative images. Thus, random selection does not guarantee the selection of the most representative and informative images to the anchor and can result in the construction of so-called *trivial triplets* (see Section II for details). We would like to note that one can also exploit all the images in the minibatch to construct triplets, as suggested in [17]. However, this choice significantly increases the total number of triplets and, thus, the computational complexity of the training phase of the retrieval system [18], [19].

To overcome the limitation of random selection, the DML methods that evaluate the hardness of images during the sampling process are introduced in the computer vision (CV) literature (see Section II for details). According to our knowledge, most of the triplet sampling methods in CV assume that each image is annotated by a single label associated with the most significant content of the considered image and, thus, rely on single-label image annotations to decide which images are positive or negative for a given anchor image. However, RS images typically consist of multiple classes and, thus, can simultaneously be associated with different class labels (i.e., multilabels). From the DML perspective, the selection of triplets from training images annotated by multilabels is more complex than that from training images labeled by single labels. To achieve accurate DML in multilabel RS CBIR, methods that accurately select a set of triplets from multilabel training images are needed.

To address this problem, we propose a novel triplet sampling method in the framework of DML designed for multilabel RS CBIR problems. Unlike the existing triplet sampling methods, the proposed method aims to select a small set of triplets from each minibatch of multilabel training images. To this end, the proposed method consists of two consecutive steps. In the first step, a small number of diverse anchors are selected based on a simple but efficient iterative algorithm. In the second step, relevant, hard and diverse positive and negative images with respect to each anchor are chosen based on a novel strategy. Then, the triplets are constructed from the selected anchors and their respective positive and negative images. Based on these consecutive steps, the proposed method constructs a small number of the most informative and representative triplets to drive DML, resulting in an accurate CBIR and also in a reduced training complexity for the considered DNN. It is worth noting that the proposed triplet sampling method is independent of the considered DNN architecture and, therefore, can be used within any DNN presented in the literature. In the experiments, different DNN architectures are considered, while the k-NN algorithm is used for the retrieval process after the characterization of the image descriptors through the considered method. Experiments carried out on two multilabel RS benchmark archives demonstrate the effectiveness of the proposed method.

The rest of the article is organized as follows. Section II presents the related works on triplet sampling. Section III introduces the proposed method. Section IV describes the considered datasets and the experimental setup, while Section V provides the experimental results. Section VI concludes our article.

II. Related Works on Triplet Sampling

The development of DML methods that aim to learn a metric space (in which semantically similar images are close to each other) is important for an accurate CBIR. It has been shown that the triplet-based DML methods perform considerably well for the CBIR tasks [16], [20]. The triplet-based DML methods use triplets of images to learn a metric space by means of the triplet loss [14]. The optimization objective

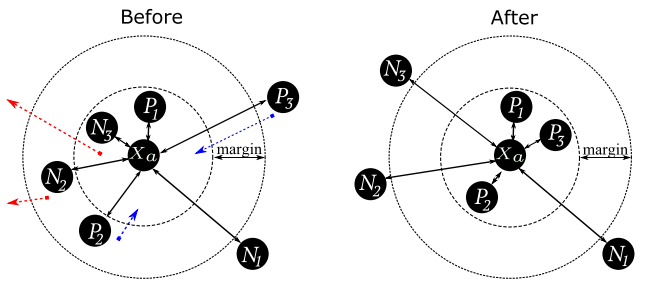

Fig. 2. Abstract representation of triplet selection and the progress for feature space update. Blue arrows indicate reducing distances for updating the embedding, while red arrows indicate increasing the distances. X_a marks a chosen anchor; P_1, P_2, and P_3 are positive images; and N_1, N_2, and N_3 are negative images in different triplets. The triplet (X_a, P_1, N_1) is trivial because it already satisfies the margins, and thus, the corresponding distances are not updated. The triplet (X_a, P_2, N_2) leads to a relatively small error, and the images are pushed and pulled a little. The triplet (X_a, P_3, N_3) violates the margin greatly and causes a significant error. P_3 is a positive image but very far from the anchor, so it is considered as a hard positive image. N_3 is, respectively, a hard negative image.

is to minimize the feature distance between the anchor and its positive sample (i.e., image) while maximizing the feature distance between the anchor and the negative sample. The goal is to ensure that the positive sample is closer to the anchor than the negative sample by at least a margin. During the training of a triplet-based DML method, for the triplets that consist of a positive image inside the margin and the negative image outside the margin, a zero value triplet loss is obtained, leading to small gradient values and slow convergence. For the triplets that consist of a positive image visually less similar to the anchor (i.e., outside the margin) and a negative image visually more similar to the anchor (i.e., inside the margin), a high triplet loss value is obtained. High loss values lead to large gradient values, and thus, the parameters of the model are updated. When a positive image is far from the margin, it is called a *hard positive* image. A negative image is called *hard negative* if it is inside the margin and very close to the anchor. If the distance between the anchor and positive image of a triplet is higher than the distance between the anchor and negative image, the triplet is considered as a *hard triplet*. In Fig. 2, an abstract representation of the triplet selection and the feature space update is demonstrated. The images P_1, P_2, and P_3 are the positive images for the anchor X_a in different triplets, while images N_1, N_2, and N_3 are the negative images for the anchor X_a. After updating the embedding (metric) space using the selected triplets, P_2 and P_3 are pulled closer to the anchor X_a, while N_2 and N_3 pushed far away from the anchor X_a toward outside the margin. The positive image P_1 is inside the margin, and the negative image N_1 is outside the margin; thus, triplet (X_a, P_1, N_1) is a trivial triplet. The positive image P_3 is a hard positive image for anchor X_a since it is outside the margin and far from the anchor image. The negative image N_3 is a hard negative image, as it is very close to the anchor. The triplet (X_a, P_3, N_3) is a hard triplet and causes a high loss value to update the parameters of the model. Since the trivial triplets are not sufficiently informative and lead to slow convergence, the use of hard triplets has been considered to overcome this problem.

Most of the methods in RS do not consider the hardness of the images in the selected triplets and exploit the random triplet selection strategy as mentioned in the introduction [16], [17], [21]. Unlike RS, in the CV community, the use of triplets is more extended, and the importance of the hardness is widely studied [22]–[25]. As an example, Xuan *et al.* [22] propose a triplet selection strategy that selects the closest positive sample (easy positive) and the closest negative (hard negative) for each anchor. Yuan *et al.* [25] propose a hard-aware deeply cascaded (HDC) embedding method. For each anchor and a selected positive sample, HDC selects the negative samples at multiple hardness levels to construct different triplets. Hardness levels are defined based on the distances in the embedding space. Yang *et al.* [19] investigate the importance of hard positive images by combining a positive image with all negative image pairs in the batch. Then, the positive images are weighted and hard positives are preferred. Ge *et al.* [24] propose a hard triplet selection method that constructs a class-level hierarchical tree of image features for the whole dataset, where visually similar classes are merged recursively. Then, the selection of the triplets is done based on a distance computed between an anchor image and different pairs of image classes through the hierarchical tree. In addition to the methods that aim to select triplets, there are also several works that focus on reformulating the triplet loss function to emphasize the effect of hard triplets [26]–[28]. As an example, Zhang *et al.* [27] adapt the focal loss that is initially defined for classification problems and propose an extended version for triplets as an alternative to the triplet loss. This loss function ensures that more importance is given to hard triplets than easier ones, and thus, the model can learn from the most informative triplets and converge faster. Kim *et al.* [26] developed an adapted version of the triplet loss for pose estimation. This loss function preserves the distance ratios from the label space in the embedding space. In [28], the multisimilarity loss function is proposed to reformulate the triplet loss with a weighting strategy. By using the weighting strategy, this loss function considers the relative similarity of all positive and all negative samples in a minibatch. In [29], the multiclass N-pair loss function is proposed to generalize the triplet loss function for multiple negative images associated with an anchor. In detail, for each anchor image, one positive image and several negative images are selected as hard negatives from different negative classes. In [21], the dual-anchor triplet loss function is introduced as an extension of the triplet loss. In addition to the objectives of the triplet loss, this loss function also aims at increasing the distance between the positive and negative images for a given anchor. Wang *et al.* [30] extend the concept of triplets to the whole minibatch, where all available images are first sorted and then divided into a positive set and a negative set. Afterward, an extension of the triplet loss is used to force a margin between the two sets by using all the images. This loss function employs a weighting strategy to increase the importance of the hard negative images. In [31], it is shown that, when an accurate sampling strategy is considered, deep learning (DL)

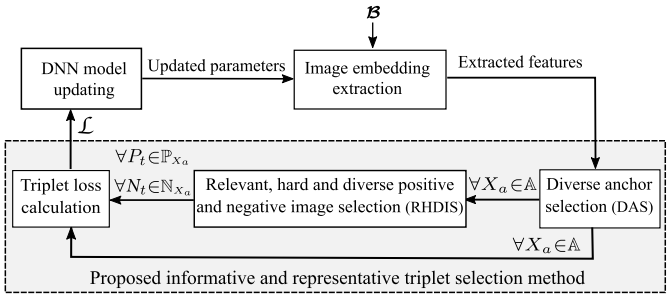

Fig. 3. Block scheme of the proposed triplet sampling method to drive the training phase of a DNN for multilabel CBIR problems.

models with different modified loss functions provide similar accuracies. This proves the fact that triplet selection is as important as loss function in the framework of DML. Most of the triplet-based methods in CV assume that a single label is associated with each image. However, RS images typically consist of multiple classes and are associated with multilabel, which makes selecting triplets more complex than the single-label scenario.

III. Proposed Method

A. Problem Formulation

Let $\boldsymbol{\mathcal{X}} = \{X_1, \ldots, X_M\}$ be an archive consisting of M images, where X_m is the mth image in the archive. We assume that a training set $\boldsymbol{\mathcal{X}}_T \subset \boldsymbol{\mathcal{X}}$ is available. Each image in $\boldsymbol{\mathcal{X}}_T$ is annotated with a set of class labels, which describes the content of the image. Let $\mathbb{L} = \{1, 2, \ldots, N\}$ be the set of all possible class labels. Each image $X_j \in \boldsymbol{\mathcal{X}}_T$ is associated with a multilabel vector $L_j = \{l_j^1, l_j^2, \ldots, l_j^N\}$, where $l_j^i = 1$ if the class label $i \in \mathbb{L}$ is associated with the image X_j, and $l_j^i = 0$ otherwise. Each training image X_j is annotated with at least one class label.

We propose a novel triplet sampling method in the framework of DL-based multilabel CBIR. The proposed method aims: 1) to select a small set of informative and representative triplets from each training minibatch $\boldsymbol{\mathcal{B}}$ and 2) to accurately describe the complex semantic content of RS images. To this end, it consists of two consecutive steps: 1) selection of anchors that are diverse to each other in the feature space and 2) selection of positive and negative images with respect to each selected anchor. To achieve the latter step, we jointly evaluate the relevancy, hardness, and diversity of the images during the selection (see Fig. 3). The proposed method is independent of the considered DL model and can be used with any DL model designed for CBIR problems. In Sections III-B and III-C, the two steps of the proposed method are described in detail.

B. Diverse Anchor Selection

The first step of the proposed method aims to find a small set of the most representative anchors. As mentioned before, all samples (i.e., images) in the minibatch $\boldsymbol{\mathcal{B}}$ could be selected as anchors. However, such an approach results in a large and redundant set of triplets and increases the computational complexity of the training. In detail, the complexity of the training grows cubically if all possible triplets are exploited [27]. Selecting a small set of anchors can significantly reduce the computational complexity of the training. To this end, we introduce a simple but efficient diverse anchor selection (DAS) strategy. The DAS strategy aims to select diverse anchors from the minibatch that, when included in the set of triplets, can improve the retrieval performance. To this end, it exhibits an iterative algorithm to evaluate the diversity in the feature space among the samples from the minibatch. The algorithm starts with an empty set $\mathbb{A} = \emptyset$. The first anchor is selected randomly from the current minibatch $\boldsymbol{\mathcal{B}}$ and added into $\mathbb{A}$. At each iteration, a new anchor that is associated with the highest distance from all already selected anchors is selected from $\boldsymbol{\mathcal{B}}$. In detail, at the hth iteration, the hth anchor image X_h is selected as

$$X_h = \underset{X_b \in \boldsymbol{\mathcal{B}} \setminus \mathbb{A}}{\operatorname{argmax}} \left[\max_{X_a \in \mathbb{A}} D(X_b, X_a) \right] \tag{1}$$

where $D(\cdot, \cdot)$ is the feature similarity measure, defined as the Euclidean distance between two images in the feature space. It is worth noting that the Euclidean distances are normalized based on min–max normalization. The steps are iterated until H anchors are selected. Due to the selection of anchors that are as distant as possible to each other in the feature space, the diversity among the selected anchors with respect to their correlation in the feature space is maximized. This results in selecting a representative set of anchors, forming the basis for the positive and negative image selection steps.

C. Relevant, Hard, and Diverse Positive–Negative Image Selection

The second step of the proposed method aims to select, for each anchor, positive and negative images that are informative (i.e., relevant and hard) and representative (i.e., diverse to each other in the feature space). This is achieved by a novel relevant, hard, and diverse positive and negative image selection strategy (RHDIS). The relevancy of an image to an anchor is defined based on its multilabel similarity with respect to the considered anchor. In detail, a positive image can be associated with high relevancy to an anchor if their class label similarity is high and vice versa. A negative image can be relevant to an anchor if its class label similarity is small and vice versa. The hardness of an image is associated with its distance to the considered anchor in the feature space. In detail, a positive image can be hard if its distance to the anchor in the embedding space is high, whereas a negative image can be considered hard if its distance to the anchor is small.

The proposed RHDIS strategy initially evaluates the informativeness (i.e., relevancy and hardness) of the images to select the candidates for positive and negative images related to each anchor image. Then, the representative (diverse) ones among the most informative positive and negative images are selected to construct the triplets. To this end, for each image X_b in the minibatch $\boldsymbol{\mathcal{B}}$, informativeness scores $I_p(X_a, X_b)$ (which shows if X_b is a candidate positive image) and $I_n(X_a, X_b)$ (which shows if X_b is a candidate negative image)

with respect to anchor X_a are initially computed as

$$I_p(X_a, X_b) = \beta \times S(X_a, X_b) + (1-\beta) \times D(X_a, X_b) \quad (2)$$

$$I_n(X_a, X_b) = \beta \times [1 - S(X_a, X_b)] + (1-\beta) \times [1 - D(X_a, X_b)] \quad (3)$$

where $S(X_a, X_b)$ shows the class label similarity between the image X_b and X_a. $S(X_a, X_b) \in [0, 1]$ is calculated based on the soft pairwise similarity measure (i.e., the distance between the multilabel vector L_a of X_a and L_b of X_b) [32]. If $S(X_a, X_b)$ is high, X_b can be considered as a relevant positive image, whereas, if $[1-S(X_a, X_b)]$ is high, X_b can be considered as a relevant negative image. $D(X_a, X_b)$ is the distance between X_b and X_a in the embedding space and measures the hardness of images as mentioned before. If both $D(X_a, X_b)$ and $S(X_a, X_b)$ are high, the image X_b can be considered as a relevant and hard positive image. If both $[1-S(X_a, X_b)]$ and $[1-D(X_a, X_b)]$ are high, the image X_b can be considered as a relevant and hard negative image. $\beta \in [0, 1]$ is the weighting parameter and can be adjusted to give more importance to either the relevancy or the hardness of the image.

To construct a set $\mathbb{P}_{X_a} = \{P_1, P_2, \ldots, P_C\}$ of C positive images for an anchor X_a, the image in the minibatch associated with the highest I_p score with respect to X_a is chosen as the first positive image. Then, the next images are iteratively selected. We apply an iterative approach similar to the DAS introduced in the first step to select the most representative images. At the tth iteration, the tth positive image P_t is selected as

$$P_t = \underset{X_b \in \boldsymbol{\mathcal{B}} \setminus \mathbb{P}_{X_a}}{\operatorname{argmax}} \left[\gamma \times I_p(X_a, X_b) + (1-\gamma) \times \max_{P_c \in \mathbb{P}_{X_a}} D(X_b, P_c) \right]. \quad (4)$$

This process is repeated until the desired number of positive images is selected. The parameter $\gamma \in [0, 1]$ controls the influence of the diversity term.

To construct a set $\mathbb{N}_{X_a} = \{N_1, N_2, \ldots, N_C\}$ of C negative images for each anchor X_a, the image with the highest I_n score in the minibatch with regard to X_a is selected as the first negative image. Afterward, the subsequent negative images are iteratively selected. At the tth iteration, the tth negative image N_t is selected as

$$N_t = \underset{X_b \in \boldsymbol{\mathcal{B}} \setminus \mathbb{N}_{X_a}}{\operatorname{argmax}} \left[\gamma \times I_n(X_a, X_b) + (1-\gamma) \times \max_{N_c \in \mathbb{N}_{X_a}} D(X_b, N_c) \right]. \quad (5)$$

This selection strategy ensures that the selected positive and negative images for each anchor are informative (i.e., hard and relevant) and representative (i.e., diverse among each other in the feature space). After selecting the final set of triplets from the minibatch $\boldsymbol{\mathcal{B}}$, the triplet loss function is calculated as

$$\mathcal{L} = \sum_{\substack{\forall X_a \in \mathbb{A} \\ \forall P_t \in \mathbb{P}_{X_a} \\ \forall N_t \in \mathbb{N}_{X_a}}} \max \left([D(X_a, P_t) - D(X_a, N_t) + \alpha], 0 \right) \quad (6)$$

where α is a margin enforced between positive and negative images for an anchor image. After an end-to-end training of the whole neural network by minimizing the triplet loss and learning the network parameters, the descriptors (i.e., features) of the images in $\boldsymbol{\mathcal{X}} \setminus \boldsymbol{\mathcal{X}}_T$ are obtained. Then, the k most semantically similar images with regard to a given query image $X_q \in \boldsymbol{\mathcal{X}}$ are selected by comparing their descriptors based on the k-NN algorithm.

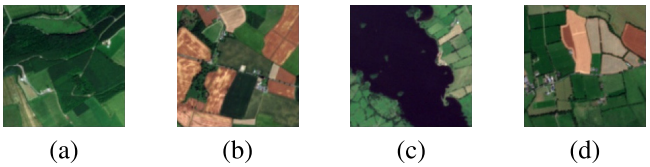

Fig. 4. Example of images from the IRS-BigEarthNet archive and their multilabels: (a) *arable land, pastures*, and *coniferous forest*; (b) *arable land* and *pastures*; (c) *pastures* and *inland waters*; and (d) *arable land, pastures*, and *complex cultivation patterns*.

Fig. 5. Example of images from the UCMerced Land Use archive and the multilabels associated with them: (a) *sand* and *sea*; (b) *airplane, cars, grass*, and *pavement*; (c) *bare soil, buildings*, and *grass*; and (d) *buildings, cars, pavement*, and *trees.*

IV. Dataset Description and Design of Experiments

A. Dataset Description

To evaluate the proposed method, we conducted experiments on two different multilabel RS archives. The first archive is BigEarthNet [15], [33], which is a large-scale multilabel Sentinel-2 benchmark archive consisting of 590 326 images. In the experiments, we considered the images acquired over Ireland in the summer of 2017 (denoted as IRS-BigEarthNet). IRS-BigEarthNet contains 15 894 images, each of which is made up of 120 × 120 pixels for 10-m bands, 60 × 60 pixels for 20-m bands, and 20 × 20 pixels for 60-m bands. In the experiments, we excluded the 60-m bands and applied bicubic interpolation to 20-m bands, which results in ten bands, each of which has a size of 120 × 120 pixels. The class labels of the images were obtained from the CORINE Land Cover database of the year 2018 (CLC 2018). In the experiments, we used the 19 class nomenclature presented in [33]. As suggested in [33], images with snow cover, cloud cover, and cloud shadows are excluded from training and evaluation. Fig. 4 shows an example of images from the IRS-BigEarthNet with the associated multilabel annotations.

The second benchmark archive is the UC Merced Land Use (UCMerced) archive [34], which consists of 2100 images selected from aerial orthoimagery with a spatial resolution of 30 cm. Each image has a size of 256 × 256 pixels. The images are annotated with multilabels by Chaudhuri *et al.* [5]. There are 17 classes in total with at least one and a maximum of seven class labels per image. Fig. 5 shows an exam-

ple of images from this archive along with their multilabel annotations.

The two benchmark archives differ greatly in size, complexity, and characteristics. This allows us to demonstrate the general applicability and success of the proposed triplet sampling method in different scenarios. We randomly split UCMerced images into 60% for training, 20% for validation, and 20% for testing. For IRS-BigEarthNet, the officially provided splits into training, validation, and evaluation sets were used. During the training step, all triplets were sampled from the training set. Query images were taken from the validation set, while image retrieval is applied to the evaluation set.

B. Design of Experiments

In the experiments, different CNN architectures were considered as backbones, while an additional fully connected layer was added to produce image embeddings. The resulting CNNs were trained for image retrieval by means of the triplet loss. It is worth noting that our method does not depend on a specific DL model architecture. In our experiments, we evaluated three different CNN architectures: 1) the shallow CNN (S-CNN) [15]; 2) DenseNet-121 [35]; and 3) ResNet-50 [36]. The last two architectures are well-known deep models, while the first architecture is an explicitly shallow model. All models were used without pretraining. The size of the minibatch for IRS-BigEarthNet and UCMerced was selected as 300 and 100, respectively. The training was performed for 100 epochs with the Adam optimizer, using an initial learning rate of 0.001 (which was exponentially decayed every five epochs by 5%). The margin parameter α of the triplet loss was set to 0.2. The values of β and γ were set to 0.5 and 0.1, respectively, based on a grid search strategy. All the experiments were conducted on NVIDIA Tesla V100 GPUs with 32-GB memory. The results were provided in terms of the different evaluation metrics, such as accuracy, precision, recall, and F_1 score [5]. These values were the average of the values obtained by retrieving the 30 and ten most similar images for IRS-BigEarthNet and UCMerced, respectively.

We carried out different kinds of experiments in order to: 1) perform a sensitivity analysis with respect to different network architectures and embedding sizes; 2) conduct an ablation study of the proposed triplet sampling method; 3) compare our method with different triplet sampling methods; and 4) compare our method with state-of-the-art DML based methods. To perform the ablation study, we compared the proposed DAS strategy (see Section III-B for the details) with two frequently used anchor selection strategies that are given as follows.

1) Batch anchor selection (BAS): This strategy selects each image in the minibatch as an anchor once and can be considered an upper bound strategy for the triplet selection. This strategy does not miss any information provided by specific triplets. However, it leads to a very high number of final triplets that can be redundant.
2) Random anchor selection (RAS): This strategy selects a fixed number of anchors from the minibatch without any prior assumption. It is simple, but there is no guarantee that the randomly chosen anchors provide a good basis for the triplets.

In the experiments, 10% of all possible anchors from the minibatch were chosen for the RAS and the proposed DAS strategies. We compared the proposed relevant, hard, and diverse positive–negative image selection (RHDIS) strategy (see Section III-C for the details) with two baselines that are:

1) Batch positive and negative image selection (BIS): This strategy uses all images in the minibatch. Each image is used as the positive and the negative images once. It covers all possible triplets, leading to a very high number of final triplets.
2) Random positive and negative image selection (RIS): This strategy randomly selects sets of positive and negative images and combines all of them into triplets. Many of the resulting triplets may be trivial, but it requires no prior knowledge and provides a lower bound baseline.

In the experiments, we also assessed the effectiveness of the joint use of the abovementioned strategies with proposed DAS and RHDIS for the selection of anchors, as well as positive and negative images. This is important as the anchor selection step is independent of the step of the positive and negative image selection, and thus, the proposed selection strategies can be combined with the other well-known strategies.

In the experiments, we also compared the proposed DAS-RHDIS method with two triplet sampling methods: 1) the DML using triplet network that uses RAS for anchor selection and RIS for positive and negative image selection (denoted as TNDML) [37] and 2) enhancing RS image retrieval using a triplet DML network, which employs BAS for the anchor selection and BIS for positive and negative image selections (denoted as RSDML) [17]. We also compared the proposed DAS-RHDIS method with state-of-the-art DML methods for CBIR: 1) the content-based medical image retrieval (CBMIR) system, which utilizes a pairwise similarity loss function to force all positive images to be close, while separating all the negative images with a fixed distance [38]; 2) the multisimilarity loss with general pair weighting for deep metric learning (denoted as MSL) [28]; 3) the dual-anchor triplet loss (denoted as DATL) proposed in [21]; and 4) the improved DML with multiclass N-pair loss objective (denoted as NPL) [29]. For all the methods, we used the same CNN architecture and training setup as in our method.

V. Experimental Results

A. Sensitivity Analysis of the Proposed Method

In this subsection, we present the results of the sensitivity analysis for the proposed triplet sampling method (denoted as DAS-RHDIS) in terms of different DL model architectures and different embedding sizes. To analyze the proposed DAS-RHDIS method in the framework of different DL models designed for multilabel RS CBIR, we selected the CNN architectures of: 1) S-CNN; 2) DenseNet-121; and 3) ResNet-50. The embedding size for each architecture was set to 256.

TABLE I

PERFORMANCE OF DIFFERENT DL MODEL ARCHITECTURES FOR THE UCMERCED ARCHIVE

Architecture	Metric (%)			
	Accuracy	Precision	Recall	F_1 Score
S-CNN	40.5	48.9	51.9	50.3
DenseNet-121	45.5	54.4	58.0	56.1
ResNet-50	**54.5**	**63.3**	**66.5**	**64.8**

TABLE II

EFFECT OF VARYING EMBEDDING SIZES ON THE RETRIEVAL PERFORMANCE FOR THE UCMERCED ARCHIVE

Embedding Size	Metric (%)			
	Accuracy	Precision	Recall	F_1 Score
256	54.5	63.3	66.5	64.8
512	56.2	64.6	69.0	66.7
1024	**56.8**	**65.3**	**70.0**	**67.5**
2048	50.3	58.4	62.8	60.5

In Table I, the results are shown for the UCMerced archive. By assessing the table, one can observe that all the considered DL model architectures provide a high performance. As an example, although S-CNN is an explicitly shallow architecture, it achieves more than 50% F_1 score as in Dense-Net-121 and ResNet-50. This shows that the proposed DAS-RHDIS method is architecture-independent. One can also see from the table that the best scores under all metrics were obtained when ResNet-50 was utilized. As an example, ResNet-50 provides almost 9% higher precision and 8.5% higher recall compared to DenseNet-121. Compared with S-CNN, ResNet-50 leads to more than 14% higher F_1 score and accuracy. These results show that a proper selection of a DL model architecture can improve performance. For the rest of the experiments, we provided the results obtained with ResNet-50 due to its proven success.

In Table II, the results obtained by using different embedding sizes are shown for the UCMerced archive. We evaluated the effect of the embedding sizes of 256, 512, 1024, and 2048 used in the proposed DAS-RHDIS method. From the table, one can see that the highest scores under all metrics are obtained when the embedding size is 1024. Further increase in the embedding size to 2048 does not improve the performance. As an example, the proposed method with the embedding size of 1024 provides a 7% higher F_1 score compared to that of 2048. This is in line with the works in the literature, which demonstrate that, beyond a certain size, adding any new embedding dimension may not improve the performance [39]–[41]. By analyzing the table, one can also observe that the lowest performance is obtained when the embedding size is 256. In this case, the F_1 score is reduced by almost 3% compared to the embedding size of 1024. Accordingly, for the rest of the experiments, we set the embedding size to 1024. These results were also confirmed through experiments obtained by using the IRS-BigEarthNet archive (not reported for space constraints).

TABLE III

RESULTS OBTAINED BY THE DIFFERENT ANCHOR SELECTION STRATEGIES (RAS, BAS, AND PROPOSED DAS) UNDER DIFFERENT METRICS FOR THE UCMERCED ARCHIVE WHEN THE PROPOSED RHDIS IS USED FOR POSITIVE AND NEGATIVE IMAGE SELECTION

Anchor Selection Strategy	Metric (%)			
	Accuracy	Precision	Recall	F_1 Score
RAS	49.2	58.1	61.9	60.0
BAS	53.5	62.0	66.5	64.2
Proposed DAS	**56.8**	**65.3**	**70.0**	**67.5**

TABLE IV

RESULTS OBTAINED BY THE DIFFERENT POSITIVE AND NEGATIVE IMAGE SELECTION STRATEGIES (RIS, BIS, AND PROPOSED RHDIS) UNDER DIFFERENT METRICS FOR THE UCMERCED ARCHIVE WHEN THE PROPOSED DAS IS USED FOR ANCHOR SELECTION

Positive and Negative Selection Strategy	Metric (%)			
	Accuracy	Precision	Recall	F_1 Score
RIS	48.6	57.4	60.1	58.7
BIS	48.9	57.6	61.4	59.4
Proposed RHDIS	**56.8**	**65.3**	**70.0**	**67.5**

B. Ablation Study

In this subsection, we performed an ablation study to analyze the effectiveness of the proposed DAS and RHDIS strategies. To demonstrate the effectiveness of the proposed DAS strategy, we compare it with RAS and BAS strategies. Table III shows the results associated with the different anchor strategies for the UCMerced archive when the proposed RHDIS strategy is used for positive and negative image selection. By analyzing the table, one can observe that the proposed DAS strategy provides the highest scores under all the metrics compared to RAS and BAS. As an example, the proposed DAS strategy provides more than 7% higher accuracy compared to RAS under the same number of anchors (which is set to ten in the experiments) when the positive and negative selection strategy is set to proposed RHDIS. In addition, the proposed DAS strategy leads to almost 4% higher recall with a smaller number of anchors compared to BAS. It is worth noting that BAS uses all the possible anchors from the minibatch (i.e., 100 anchors). This shows the success of the proposed DAS strategy to select diverse and representative anchors with respect to random sampling and batch selection strategies.

In order to demonstrate the effectiveness of the proposed RHDIS strategy, we compare it with RIS and BIS strategies. Table IV shows the results associated with the different positive and negative image selection strategies for the UCMerced archive when the proposed DAS strategy is used for anchor selection. From the table, one can see that the proposed RHDIS strategy achieves the highest performance under all metrics compared to RIS and BIS. As an example, the recall of the proposed RHDIS strategy is more than 8% higher compared to that of BIS when the anchor selection strategy is set to the proposed DAS. It is worth noting that BIS exploits all positive

TABLE V
PERFORMANCE OF DIFFERENT TRIPLET SELECTION METHODS FOR THE IRS-BIGEARTHNET AND UCMERCED ARCHIVES

Archive	Method	Metric (%)			
		Accuracy	Precision	Recall	F_1 Score
IRS-BigEarthNet	TNDML [37]	59.3	73.7	73.8	73.8
	RSDML [17]	60.2	75.4	73.9	74.6
	Proposed DAS-RHDIS	**62.7**	**77.7**	**75.7**	**76.7**
UCMerced	TNDML [37]	44.0	52.6	55.8	54.2
	RSDML [17]	48.4	56.3	61.9	59.0
	Proposed DAS-RHDIS	**56.8**	**65.3**	**70.0**	**67.5**

and negative images in the batch, while RHDIS relies on a much smaller number of triplets to achieve this result. The performance of RIS is lower than RHDIS and BIS under each metric when the anchor selection strategy is set to the proposed DAS. For example, the recall obtained by RIS is about 10% lower than that of the proposed RHDIS under the same number of triplets. This shows the effectiveness of the proposed RHDIS selection strategy to select relevant, hard, and diverse positive–negative images compared to random sampling and batch selection strategies for a given set of anchors. These results were also confirmed through experiments obtained by using the IRS-BigEarthNet archive.

C. Comparison of the Proposed Method With Different Triplet Sampling Methods

In this subsection, we evaluate the effectiveness of the proposed DAS-RHDIS method compared to different triplet selection methods, which are TNDML [37], and RSDML [17]. Table V shows the corresponding image retrieval performances on the IRS-BigEarthNet and the UCMerced archives. By analyzing the table, one can see that the proposed DAS-RHDIS method leads to the highest scores under all metrics for both archives. For example, DAS-RHDIS outperforms TNDML by 4% in precision and more than 3% in accuracy for the IRS-BigEarthNet archive, and more than 13% in F_1 score and almost 15% in recall for the UCMerced archive. The proposed DAS-RHDIS method provides about 2% higher and 8% higher F_1 scores compared to the RSDML method for IRS-BigEarthNet and UCMerced, respectively. These results demonstrate the success of the proposed DAS-RHDIS method compared to other triplet sampling methods.

Fig. 6 shows an example of images retrieved from IRS-BigEarthNet by TNDML, RSDML, and the proposed DAS-RHDIS when the query image contains *arable land*, *pastures*, and *complex cultivation patterns*. The retrieval order of images is given below the query image. By analyzing the figure, one can observe that the classes of *pasture* and *arable land* are very prominent in all retrieved images by RSDML and DAS-RHDIS, while TNDML provides similar images to the query only at the retrieved orders of five and ten. When DAS-RHDIS is compared with RSDML, the proposed method retrieves semantically more similar images. One of the reasons is that the RSDML relies only on the class label similarity, while the proposed DAS-RHDIS method: 1) extracts and exploits the semantic content of the images and

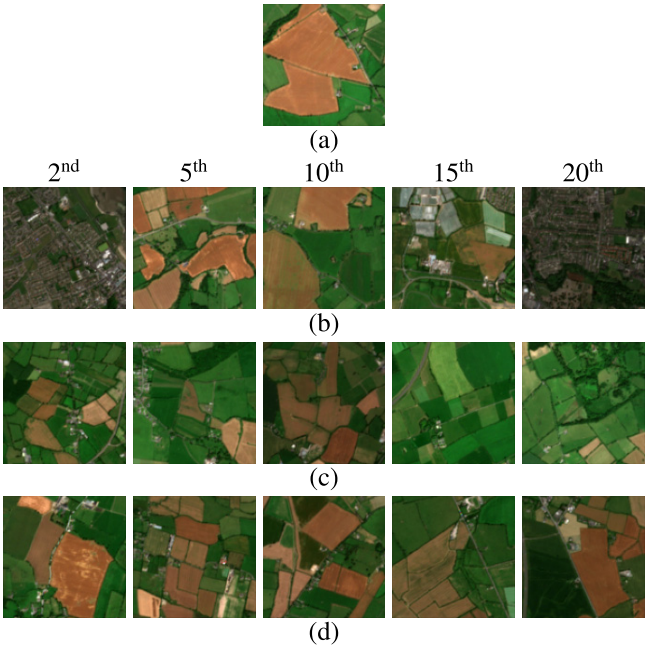

Fig. 6. Image retrieval example: (a) query image; (b) images retrieved by TNDML; (c) images retrieved by RSDML; and (d) images retrieved by the proposed DAS-RHDIS method (IRS-BigEarthNet archive).

2) considers the diversity and hardness of images during triplet selection. We observed similar behavior for the UCMerced archive. Fig. 7 shows an example of images retrieved from UCMerced. The query image for this example only contains the *Field* class. Most of the images retrieved by the proposed method (except the 20th image) belong to the same class with the query [see Fig. 7(d)]. However, only a small number of images retrieved by the TNDML and the RSDML methods contain the *field* class [see Fig. 7(b) and (c)].

During the learning of a metric space by using the triplet loss, a small subset of the available triplets carries the information needed to learn an accurate representation for image retrieval. The proposed DAS-RHDIS identifies these triplets and only learns from a subset of selected informative and representative samples, reducing the number of training triplets. Fig. 8 shows the performance of TNDML, RSDML, and the proposed DAS-RHDIS method in terms of the number of accumulated training triplets under the same number of epochs (which is set to 100 in the experiments) for the UCMerced archive. The horizontal axis shows the number of triplets in a logarithmic scale, while the vertical axis shows the

TABLE VI
PERFORMANCE OF DIFFERENT DML METHODS FOR THE IRS-BIGEARTHNET AND UCMERCED ARCHIVES

Archive	Method	Metric (%)			
		Accuracy	Precision	Recall	F_1 Score
IRS-BigEarthNet	CBMIR [38]	59.6	73.2	74.6	73.9
	MSL [28]	57.9	75.0	68.7	71.7
	DATL [21]	60.6	75.3	74.0	74.7
	NPL [29]	60.8	76.5	72.6	74.5
	Proposed DAS-RHDIS	**62.7**	**77.7**	**75.7**	**76.7**
UCMerced	CBMIR [38]	42.0	50.9	53.0	51.9
	MSL [28]	46.6	58.1	61.0	59.5
	DATL [21]	48.7	57.2	60.7	58.9
	NPL [29]	51.8	61.5	58.7	60.1
	Proposed DAS-RHDIS	**56.8**	**65.3**	**70.0**	**67.5**

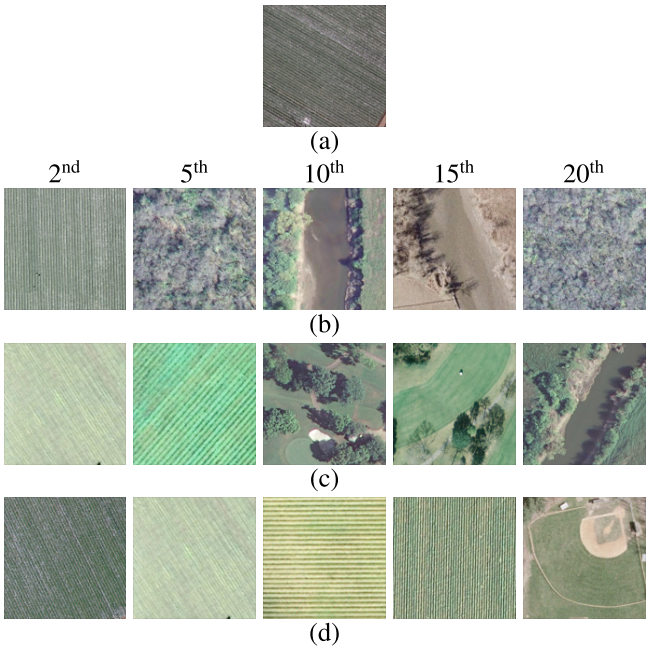

Fig. 7. Image retrieval example: (a) query image; (b) images retrieved by TNDML; (c) images retrieved by RSDML; and (d) images retrieved by the proposed DAS-RHDIS method (UCMerced archive).

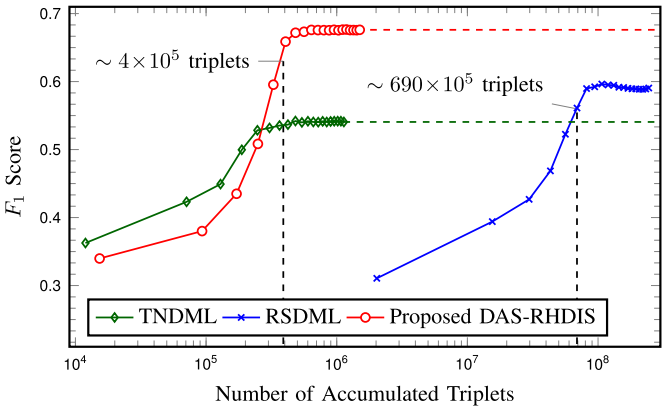

Fig. 8. F_1 scores obtained by different triplet sampling strategies and the number of accumulated triplets during the training (the UCMerced archive).

corresponding F_1 scores. The performance is associated with the numbers of triplets, which are utilized by the considered triplet selection method. The annotation points indicate the number of triplets needed for the considered method to reach at least 90% of its final performance. From the figure, one can observe that, even after the last training epoch of the proposed DAS-RHDIS method, the total number of triplets is significantly smaller than the first epoch of the RSDML method. During training, the RSDML selects more triplets at each epoch compared to the other two methods. This is due to the characteristic of RSDML that selects all the possible triplets from a minibatch, which grows cubically. The final F_1 score of our proposed method is more than 8% higher than RSDML with significantly less number of total triplets. One can also see from the figure that TNDML (which uses random triplet selection) under the same number of triplets with our method leads to a significant performance drop. The F_1 score obtained by TNDML is 13% lower than the F_1 score obtained by the proposed DAS-RHDIS method. These results show the effectiveness of our method to select a subset of informative triplets during training, resulting in faster convergence and a performance gain in the retrieval.

D. Comparison of the Proposed Method With the State-of-the-Art DML Approaches

In this subsection, we assessed the effectiveness of the proposed DAS-RHDIS method compared to the state-of-the-art DML approaches, which are CBMIR [38], MSL [28], DATL [21], and NPL [29]. Table VI shows the results under different metrics for the IRS-BigEarthNet and UCMerced archives. By analyzing the table, one can see that the proposed DAS-RHDIS method leads to the highest scores under all metrics for both archives. As an example, the proposed DAS-RHDIS method provides 2% higher and 8% higher accuracy compared to the DATL method for IRS-BigEarthNet and UCMerced, respectively. The table also shows that the CBMIR and the MSL methods obtain the lowest scores in most of the metrics. For example, CBMIR provides more than 4% lower and 14% lower precision than the proposed DAS-RHDIS for IRS-BigEarthNet and UCMerced, respectively. Since the loss function in CBMIR forces a fixed distance for all images, it is more restrictive compared to the triplet-based DML losses. This can lead to learning the metric space, in which

the similarity between the images is not properly characterized [31]. When compared with the MSL method, DAS-RHDIS achieves 7% higher recall and more than 4% higher accuracy for the IRS-BigEarthNet archive, and more than 7% higher precision and 8% higher F_1 score for the UCMerced archive. Despite the proven success of the MSL method for single-label images, we observed that the full capacity of this method is not applicable for multilabel images. Since the MSL method considers all the possible negatives and positives, and their relative feature distances among each other, its performance is very sensitive to the proper definition of the positive and the negative sets for a given anchor. However, the evident distinction of these sets is difficult to achieve for multilabel images. When compared with the NPL method, the proposed DAS-RHDIS method provides 2% higher and 7% higher F_1 scores for IRS-BigEarthNet and UCMerced, respectively. It is worth noting that NPL obtains relatively closer results to the proposed DAS-RHDIS due to its negative mining strategy. NPL uses an extension of the triplet loss, which selects multiple negative images from different negative classes for each anchor and positive image. This negative mining strategy allows NPL to include class-based diversity among the negative samples. However, in NPL, the hardness and diversity in the positive samples are not considered, resulting in the selection of trivial triplets. This can affect its performance for the retrieval task. The proposed DAS-RHDIS identifies informative and representative triplets by relying on the relevancy, hardness, and diversity of images. This allows us to reach more effective image retrieval performance compared to the other methods.

VI. Conclusion

This article introduces a novel method to select a set of informative and representative triplets from multilabel training images to achieve DML for multilabel CBIR problems in RS. The proposed triplet sampling method is defined based on a two-steps procedure and applied to each training minibatch of a DL-based retrieval system. In the first step, diverse anchor images are selected based on a simple but efficient iterative algorithm. Then, in the second step, sets of positive and negative images for each anchor are selected based on relevancy, hardness, and diversity of the positive and negative images. Finally, the triplets are constructed from the selected anchors and their respective positive and negative images. Through the abovementioned steps, the proposed method results in selecting a compact subset of informative and representative triplets, which enables accurate and efficient learning of DL models for multilabel CBIR in RS. Experimental results obtained on two multilabel RS benchmark archives under different DL architectures show the effectiveness of the proposed method in CBIR problems. In detail, the results have demonstrated that most of the available triplets do not contribute to the learning progress and can be safely discarded. Focusing on a small informative and representative subset is sufficient for achieving comparable performance compared to the case, for which all possible triplets are used. It is worth noting that the proposed triplet sampling method does not rely on a specific DL architecture and can be adapted to any metric learning method.

As a final remark, we would like to point out that the proposed method currently relies on the class labels to select positive and negative images for each anchor. As future work, we plan to develop an unsupervised strategy that can select informative positive and negative images without requiring any land-use land-cover class label.

Acknowledgment

The authors would like to thank Tristan Kreuziger for the valuable discussions on triplet sampling and for the design of Fig. 2.

References

[1] G. Sumbul, J. Kang, and B. Demir, "Deep learning for image search and retrieval in large remote sensing archives," in *Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science and Geosciences*. Hoboken, NJ, USA: Wiley, 2021, ch. 11, pp. 150–160.

[2] Y. Yang and S. Newsam, "Geographic image retrieval using local invariant features," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 2, pp. 818–832, Feb. 2013.

[3] 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.

[4] I. Tekeste and B. Demir, "Advanced local binary patterns for remote sensing image retrieval," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2018, pp. 6855–6858.

[5] 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. 2018.

[6] B. Demir and L. Bruzzone, "Hashing-based scalable remote sensing image search and retrieval in large archives," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 2, pp. 892–904, Sep. 2016.

[7] B. Chaudhuri, B. Demir, L. Bruzzone, and S. Chaudhuri, "Region-based retrieval of remote sensing images using an unsupervised graph-theoretic approach," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 7, pp. 987–991, Jul. 2016.

[8] H. Lai, Y. Pan, Y. Liu, and S. Yan, "Simultaneous feature learning and hash coding with deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3270–3278.

[9] Y. Li, Y. Zhang, C. Tao, and H. Zhu, "Content-based high-resolution remote sensing image retrieval via unsupervised feature learning and collaborative affinity metric fusion," *Remote Sens.*, vol. 8, no. 9, p. 709, Aug. 2016.

[10] X. Tang, X. Zhang, F. Liu, and L. Jiao, "Unsupervised deep feature learning for remote sensing image retrieval," *Remote Sens.*, vol. 10, no. 8, p. 1243, Aug. 2018.

[11] R. Imbriaco, C. Sebastian, and E. Bondarev, "Aggregated deep local features for remote sensing image retrieval," *Remote Sens.*, vol. 11, no. 5, p. 493, Jan. 2019.

[12] Y. Boualleg and M. Farah, "Enhanced interactive remote sensing image retrieval with scene classification convolutional neural networks model," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2018, pp. 4748–4751.

[13] F. Sabahi, M. O. Ahmad, and M. N. S. Swamy, "An unsupervised learning based method for content-based image retrieval using Hopfield neural network," in *Proc. 2nd Int. Conf. Signal Process. Intell. Syst.*, Dec. 2016, pp. 1–5.

[14] 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.

[15] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, "BigEarthNet: A large-scale benchmark archive for remote sensing image understanding," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2019, pp. 5901–5904.

[16] S. Roy, E. Sangineto, B. Demir, and N. Sebe, "Metric-learning-based deep hashing network for content-based retrieval of remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 2, pp. 226–230, Feb. 2021.

[17] R. Cao *et al.*, "Enhancing remote sensing image retrieval using a triplet deep metric learning network," *Int. J. Remote Sens.*, vol. 41, no. 2, pp. 740–751, Jan. 2020.

[18] C. Zhou *et al.*, "Angular deep supervised hashing for image retrieval," *IEEE Access*, vol. 7, pp. 127521–127532, 2019.

[19] X. Yang, P. Zhou, and M. Wang, "Person reidentification via structural deep metric learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 10, pp. 2987–2998, Oct. 2019.

[20] P. Zhu, Y. Tan, L. Zhang, Y. Wang, and J. Mei, "Deep learning for multilabel remote sensing image annotation with dual-level semantic concepts," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 6, pp. 4047–4060, Jan. 2020.

[21] M. Zhang, Q. Cheng, F. Luo, and L. Ye, "A triplet nonlocal neural network with dual-anchor triplet loss for high-resolution remote sensing image retrieval," *IEEE J. Sel. Topics Appl. Earth Observ., Remote Sens.*, vol. 14, pp. 2711–2723, 2021.

[22] H. Xuan, A. Stylianou, and R. Pless, "Improved embeddings with easy positive triplet mining," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2474–2482.

[23] D. Zhang, Y. Li, and Z. Zhang, "Deep metric learning with spherical embedding," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 18772–18783.

[24] W. Ge, W. Huang, D. Dong, and M. R. Scott, "Deep metric learning with hierarchical triplet loss," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 269–285.

[25] Y. Yuan, K. Yang, and C. Zhang, "Hard-aware deeply cascaded embedding," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 814–823.

[26] S. Kim, M. Seo, I. Laptev, M. Cho, and S. Kwak, "Deep metric learning beyond binary supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 2283–2292.

[27] S. Zhang, Q. Zhang, X. Wei, Y. Zhang, and Y. Xia, "Person re-identification with triplet focal loss," *IEEE Access*, vol. 6, pp. 78092–78099, 2018.

[28] X. Wang, X. Han, W. Huang, D. Dong, and M. R. Scott, "Multi-similarity loss with general pair weighting for deep metric learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 5017–5025.

[29] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1857–1865.

[30] X. Wang, Y. Hua, E. Kodirov, G. Hu, R. Garnier, and N. M. Robertson, "Ranked list loss for deep metric learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 5202–5211.

[31] C.-Y. Wu, R. Manmatha, A. J. Smola, and P. Krahenbuhl, "Sampling matters in deep embedding learning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2840–2848.

[32] Z. Zhang, Q. Zou, Y. Lin, L. Chen, and S. Wang, "Improved deep hashing with soft pairwise similarity for multi-label image retrieval," *IEEE Trans. Multimedia*, vol. 22, no. 2, pp. 540–553, Feb. 2020.

[33] G. Sumbul *et al.*, "BigEarthNet-MM: A large scale multi-modal multi-label benchmark archive for remote sensing image classification and retrieval," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 3, pp. 174–180, May 2021.

[34] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. Int. Conf. Adv. Geograph. Inf. Syst.*, 2010, pp. 270–279.

[35] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2261–2269.

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

[37] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *Proc. Int. Workshop Similarity-Based Pattern Recognit.*, 2015, pp. 84–92.

[38] S. Deepak and P. M. Ameer, "Retrieval of brain MRI with tumor using contrastive loss based similarity on GoogLeNet encodings," *Comput. Biol. Med.*, vol. 125, Oct. 2020, Art. no. 103993.

[39] H. Xuan, R. Souvenir, and R. Pless, "Deep randomized ensembles for metric learning," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 723–734.

[40] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4004–4012.

[41] W. Chen *et al.*, "Deep image retrieval: A survey," 2021, *arXiv:2101.11282*.

Gencer Sumbul (Graduate Student Member, IEEE) received the B.S. and M.S. degrees in computer engineering from Bilkent University, Ankara, Turkey, in 2015 and 2018, respectively. He is currently pursuing the Ph.D. degree with the Faculty of Electrical Engineering and Computer Science, Technische Universität Berlin, Berlin, Germany.

He has been a Research Associate with the Remote Sensing Image Analysis (RSiM) Group, Faculty of Electrical Engineering and Computer Science, Technische Universität Berlin, since 2019. His research interests include computer vision, pattern recognition, and machine learning, with a special interest in deep learning, large-scale image understanding, and remote sensing.

Dr. Sumbul is a Referee for journals such as the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE ACCESS, the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, and the *ISPRS Journal of Photogrammetry and Remote Sensing*, and international conferences, such as European Conference on Computer Vision and IEEE International Geoscience and Remote Sensing Symposium.

Mahdyar Ravanbakhsh (Member, IEEE) received the Ph.D. degree in cognitive environments from the University of Genoa, Genoa, Italy, in 2019.

He was a Post-Doctoral Research Fellow with the University of Genoa. Before joining the Technische Universität Berlin, Berlin, Germany, he was a Post-Doctoral Research Fellow with the Department of Engineering and Naval architecture (DITEN), University of Genoa. In 2016, he was a Guest Researcher with the Deep Relational Learning Group, University of Trento, Trento, Italy. He has been a Research Fellow with the Remote Sensing Image Analysis (RSiM), Department of Electrical Engineering and Computer Science, Technische Universität Berlin, since 2019. His research lies at the intersection of machine learning and computer vision with an emphasis on deep learning with minimal supervision and/or limited data.

Dr. Ravanbakhsh was a recipient of the "Best Student Paper Award: Second Place" at the 2017 International Conference on Image Processing (ICIP).

Begüm Demir (Senior Member, IEEE) received the B.S., M.Sc., and Ph.D. degrees in electronic and telecommunication engineering from Kocaeli University, Kocaeli, Turkey, in 2005, 2007, and 2010, respectively.

She is currently a Full Professor and the Founder Head of the Remote Sensing Image Analysis (RSiM) Group, Faculty of Electrical Engineering and Computer Science, Technische Universität Berlin (TU Berlin), Berlin, Germany, and the Head of the Big Data Analytics for Earth Observation Research Group, Berlin Institute for the Foundations of Learning and Data (BIFOLD), Berlin. Specifically, she performs research in the field of processing and analysis of large-scale Earth observation data acquired by airborne and satellite-borne systems. In 2018, she received a Starting Grant from the European Research Council (ERC) for her project "BigEarth: Accurate and Scalable Processing of Big Data in Earth Observation." Her research activities lie at the intersection of machine learning, remote sensing, and signal processing.

Dr. Demir is a fellow of the European Laboratory for Learning and Intelligent Systems (ELLIS). She is a Scientific Committee Member of several international conferences and workshops, such as Conference on Content-Based Multimedia Indexing, Conference on Big Data from Space, Living Planet Symposium, International Joint Urban Remote Sensing Event, SPIE International Conference on Signal and Image Processing for Remote Sensing, Machine Learning for Earth Observation Workshop organized within the ECML/PKDD. She was a recipient of the prestigious "2018 Early Career Award" by the IEEE Geoscience and Remote Sensing Society for her research contributions in machine learning for information retrieval in remote sensing. She is a Referee for several journals such as the PROCEEDINGS OF THE IEEE, the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, the IEEE TRANSACTIONS ON IMAGE PROCESSING, Pattern Recognition, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, the *International Journal of Remote Sensing*, and several international conferences. She is currently an Associate Editor for the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS and *MDPI Remote Sensing and International Journal of Remote Sensing*.