Cross-modal Remote Sensing Image Retrieval

by

DHYANIL MEHTA
202011032

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of

MASTER OF TECHNOLOGY
in
INFORMATION AND COMMUNICATION TECHNOLOGY
to

DHIRUBHAI AMBANI INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGY

May, 2022

Declaration

I hereby declare that

i) the thesis comprises of my original work towards the degree of Master of Technology in Information and Communication Technology at Dhirubhai Ambani Institute of Information and Communication Technology and has not been submitted elsewhere for a degree,

ii) due acknowledgment has been made in the text to all the reference material used.

Dhyanil Mehta

Certificate

This is to certify that the thesis work entitled *Cross-modal Remote Sensing Image Retrieval* has been carried out by *Dhyanil Mehta* for the degree of Master of Technology in Information and Communication Technology at *Dhirubhai Ambani Institute of Information and Communication Technology* under our supervision.

Avik Hati
Thesis Supervisor

Biplab Banerjee
Thesis Co-Supervisor

Contents

Abstract

Most of the conventional remote sensing (RS) retrieval approaches used today are often based on a single modality data framework. In today's date, the need for multimodal and cross-modal based approaches especially in the RS retrieval area are growing evident with more and more data being acquired from different satellite sensors. This thesis presents a few-shot learning based cross-modal image retrieval framework for RS images. Few-shot learning was incorporated to account for label scarcity or when the data available is insufficient and DeepCORAL loss was further integrated for domain adaptation of the cross-modal data. In addition, a reciprocal points loss is also integrated for generating better discriminative features of images. We evaluate our approach on two cross-source remote sensing image datasets by training cross-modally and testing uni-modally on insufficient labeled data and achieve positive results showing our framework to be helpful.

List of Principal Symbols and Acronyms

A	Abundant data modality image space
L	Class label space
L_{coral}	DeepCORAL loss
L_{proto}	Prototypical cross-entropy loss
L_{rpl}	Reciprocal points loss
S	Scarce data modality image space

List of Tables

List of Figures

CHAPTER 1
Introduction

Image retrieval is a well-researched problem in the field of computer vision, where images in the database similar to a given query image are retrieved [11]. A similarity measure is used between the query and database/gallery images to rank the database images in decreasing order of similarity. First, image features obtained from a reliable feature extractor for the database images are stored beforehand. During retrieval, given a query image, its features are computed and compared with the pre-computed database image features. For this, a similarity value is obtained, and the images are ranked accordingly. Let us denote the query image as Q, an image in the gallery/database D as I, feature extractor model as f_θ, and similarity measure as S. The similarity is calculated as:

$$sim_I = S(f_\theta(Q), f_\theta(I)) \tag{1.1}$$

Figure 1.1 shows the overview of remote sensing image retrieval process.

In recent times, there have been many signs of progress in the remote sensing (RS) area. Both the quantity and quality of remote sensing images are growing rapidly. Designing an effective feature extraction method based on remote sensing images' characteristics can improve the retrieval performance. Various fruitful efforts have been made in building efficient and accurate remote sensing image retrieval methods for use in searching large remote sensing (RS) archives [23].

Early studies particularly emphasized finding various feature representation methods to improve the accuracy of RS image retrieval, aiming to find features or feature combinations that discriminates different classes well [33]. In current times, as the complexity of RS data is increasing, ambiguity has increased for the visual features [26]. Thus, basic RS retrieval systems no longer give satisfactory performance. High-level features for RS data were thus derived using convolutional neural networks (CNNs) for high-resolution remote sensing (HRRS) image retrieval to overcome this difficulty [43, 39, 44]. CNN features have proven themselves to be having a solid discrimination ability and improve retrieval perfor-

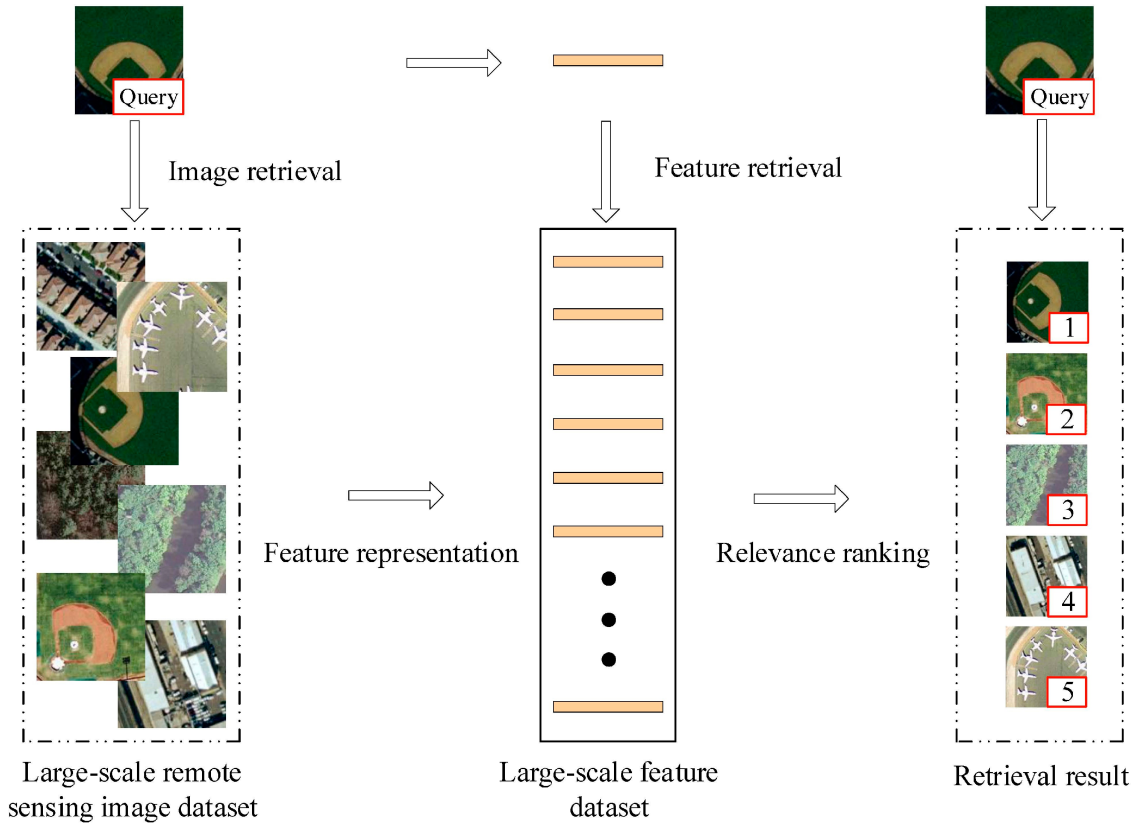


Figure 1.1: Remote sensing image retrieval [15]

mance in several computer vision tasks.

Nevertheless, most of the works focused on training feature extractors for unimodal retrieval or where only a single modality of data is used [43, 44, 39, 41]. If the training data available is insufficient, which can be the case for some RS archives, the retrieval performance also drops. Consequently, researchers proposed to use similar data of different modalities to help train reliable retrieval frameworks [19, 31, 40, 4, 22, 17]. This method to retrieve images of one modality with the help of data of different modalities is called cross-modal image retrieval. Different modalities of data such as text, audio, images from different datasets have been used in these approaches.

As we are considering insufficient training data, it makes sense to bring few-shot learning into the fold, which is rapidly gaining popularity in the deep learning community. Few-shot learning is designed for tasks whereonly a limited number of labeled samples are available, and the task is to generalize from those few examples only [36]. Some of the research works have leveraged this for solving problems in the RS domain [5, 42].

1.1 Problem Definition

In this thesis, we solve the problem of remote sensing image retrieval on databases that do not have sufficient training data by leveraging a similar remote sensing database containing co-registered or paired images and having labeled samples in abundance.

1.2 Key Contributions

The key contributions of this thesis are as follows:

- A reliable deep feature extractor for RSIR is developed for the case of insufficient training data.
- A few-shot learning based training framework is proposed.
- Cross-modality is integrated in the framework to take advantage of sufficiently available labeled RS data of a different modality by using DeepCORAL loss [30] for domain adaptation of two modalities.
- Reciprocal points loss [6] is used for better clustering of the classes and maximizing distance between them, thus enhancing the discriminative ability of the deep feature extractor.

CHAPTER 2
Uni-modal Remote Sensing Image Retrieval

The research done in the field of remote sensing image retrieval can be broadly classified into two categories namely, uni-modal and cross-modal retrieval methods. In this chapter, we discuss retrieval considering remote sensing images of a single modality i.e., uni-modal retrieval methods. Cross-modal remote sensing image retrieval methods are discussed in Chapter 3.

Remote sensing image retrieval task falls under the domain of object detection. We first talk about approaches using low-level and mid-level features for the retrieval task, and then move on to high-level features which can be leveraged using CNNs.

2.1 Low-level and Mid-level Features

The most prominent information in RS images is described by spectral characteristics, which are one of the basic features [24]. Spectral characteristics have been used to retrieve RS data in a variety of ways. They save the reflectance information of the comparable areas of the Earth's surface. This results in extreme sensitivity to noise and changes in lighting. Features like scale-invariant feature transform (SIFT) [21] have also proved effective for RSIR.

Mid-level features can be extracted by first computing local image descriptors like texture, spectral, local invariant features, and combining them into effective representation using encoding methods like BoW [27], FV [25], and VLAD [18].

BoW [27] is a frequently used basic encoding approach that builds a visual codebook using k-means clustering and counts local features in the codebook histogram. It has been used in various RS image retrieval studies and has yielded positive results [38, 2].

VLAD [18] is a more advanced variant of BoW that computes the distance between local features and cluster centres in addition to feature distribution. On HRRS images, VLAD is used to encode local pattern spectra and produce high-

precision retrieval results [3].

Next, we discuss some notable methods for the task of RSIR using high-level features.

2.2 High-level features

With the advent of deep learning, especially convolutional neural networks (CNNs), extracting useful high-level features has become easy. CNNs trained for the task of classification can produce some very discriminative high-level features which can be used for the task of image retrieval.

When a classification network is used for retrieval, feature extraction and similarity measurement are implemented independently. A CNN-based classifier, after training, is used to extract discriminative features as the representation for that image. The features are typically taken from the last convolutional layers as they are usually more representative and useful than the output layer fully-connected features.

Napoletano [23] provided an extensive analysis on visual descriptors for content-based retrieval from remote sensing images. He found out the CNN features to be performing much better than global or hand-crafted features. He further highlighted the importance of domain adaptation for remote sensing images by evaluating the performance of ResNet50 fine-tuned on RS domain image datasets.

Zhou *et al.* [43] combined a standard CNN with a three-layer perceptron layer also known as mlpconv layer and proposed a novel CNN architecture called low-dimensional CNN (LDCNN) for remote sensing image retrieval. As shown in Figure 2.1, the LDCNN architecture is composed of five linear convolution layers, an mlpconv layer and a global average pooling layer. The proposed architecture is based on the assumption that the first several convolutional layers learn linearly separable features such as edges and corners, and the last layers learn more abstract high-level features which are not linearly separable.

Gabor-CA-ResNet by Zhuo and Zhou [44] is another attempt to provide an efficient retrieval framework. This network is designed by modifying the ResNet network to get an effective representation that captures the complexity of remote sensing images. The deep features are first extracted from a modified ResNet50 network. Then a Gabor convolutional layer is added to further obtain a rich representation as Gabor has proven to be effective in describing the image space. A channel attention mechanism is further added to obtain semantic features, further enhancing the ability of deep features obtained. The authors also proposed a split-

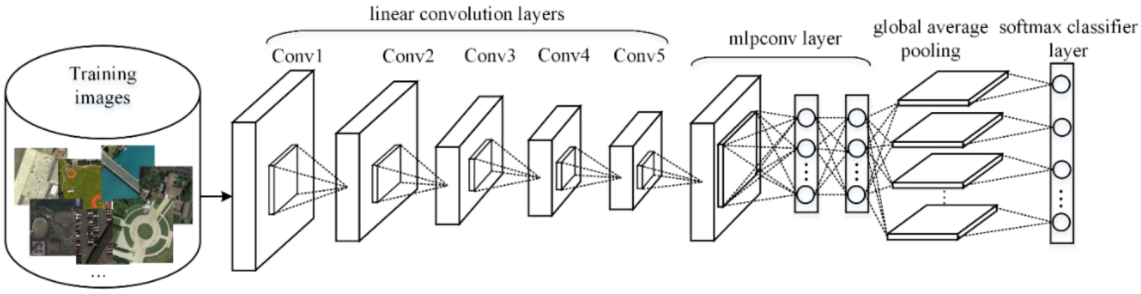


Figure 2.1: Low-dimensional CNN (LDCNN) architecture [43]

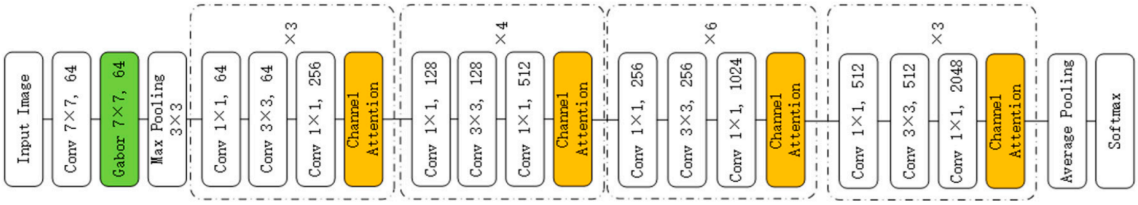


Figure 2.2: Gabor-CA-ResNet architecture [44]

based deep feature transform network to reduce the dimensionality of the deep features obtained from Gabor-CA-ResNet network. The architecture is shown in Figure 2.2.

Ye *et al.* [39] proposed a weighted distance approach to calculate the feature similarity for retrieval. Pre-trained VGG and ResNet networks are first fine-tuned on RS domain images to get relevant deep features. Thereafter while calculating the similarity between query and database image features, a weight parameter calculated from the class probability of the query image is introduced and multiplied with the similarity value. This weight gives preference to the retrieved images in similar classes with the query image. As a result, there are less irrelevant images in retrieving and ranking process. The weight of a retrieved image r belonging to class k is calculated as:

$$w = 1 - p_k^q \tag{2.1}$$

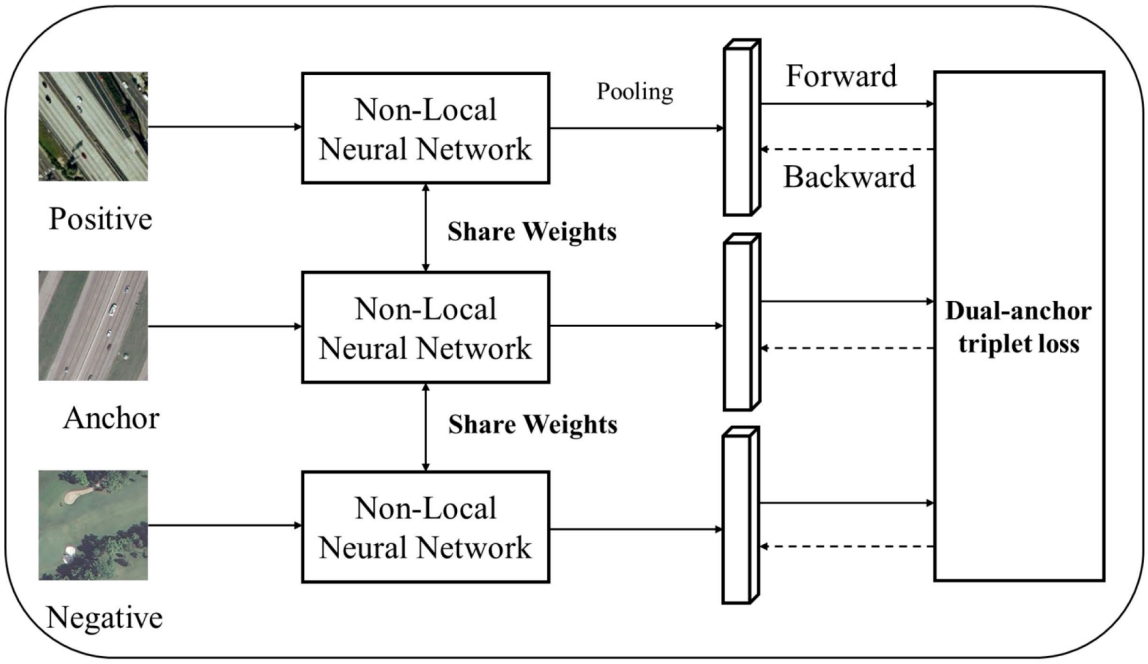
where p_k^q is the predicted probability of query image q being in the class k which is also the class of retrieved image r. The weighted distance is then obtained by the following equation,

$$d_w(q,r) = w \times d(q,r) \tag{2.2}$$

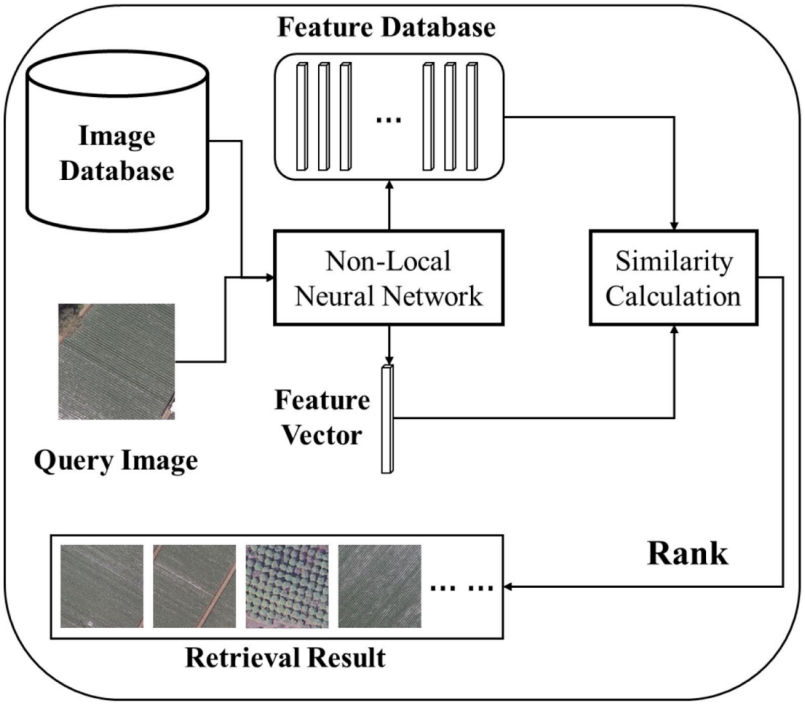
where $d(q,r)$ is the distance metric used. The authors used Euclidean distance as distance metric for evaluating their method.

Zhang *et al.* in [41] used the concept of non-local neural networks [35] for

the task and constructed a three branch network design called triplet non-local neural network (T-NLNN). As shown in the training workflow in Figure 2.3, the three-branch or triplet-loss based architecture is used to train the non-local neural network (NLNN) model. They also modified the existing triplet loss to also consider the distance between positive and negative samples, thus calling it a dual-anchor triplet loss. During the retrieval stage shown in Figure 2.3, NLNN is used as a deep feature extractor to extract the features which are then compared for similarity matching and ranking.



(a) Training stage



(b) Retrieval stage

Figure 2.3: Architecture of T-NLNN [41]

CHAPTER 3
Cross-modal Remote Sensing Image Retrieval

Cross-modal retrieval methods are useful when training data in the primary modality is not sufficiently available or a different modality data can help in improving the performance. In cross-modal image retrieval, the query data is of a different modality than the image database. We will use cross-modality approach in our framework in the training stage. Most of the works for cross-modal retrieval for RS domain can be categorized into three retrieval frameworks namely, image-image retrieval [19, 31, 40, 4], audio-image retrieval [4, 22], and text-image retrieval [17].

3.1 Image-Image Cross-modal RSIR

Hashing methods based on cross-modality are widely popular for cross-modal image retrieval. Many of the cross-modal approaches use a hashing based approach in which the images are converted into compact hash codes by non-linear hashing models/functions. After a good hashing function is learned, the original task can be converted to a hash-code based retrieval task which is much easier.

While uni-modal hashing functions are trained on same modality data, cross-modal based methods train hashing functions to transform cross-modal data into hash codes where different modalities of data are projected onto a common Hamming space. The query and gallery/database data here are of different modalities. Li *et al.* [19] used panchromatic and multispectral images, and proposed a new source-invariant deep hashing convolutional neural networks (SIDHCNNs). It can be optimized from scratch and used in an end-to-end manner. They also contributed a dual source remote sensing image dataset (DSRSID) consisting of panchromatic and multispectral images to the community.

Sun *et al.* [31] proposed a semantic preserving deep hashing method for cross-modal RSIR and created a new cross-domain remote sensing dataset of very high resolution (VHR) and synthetic aperture radar (SAR) images. The method con-

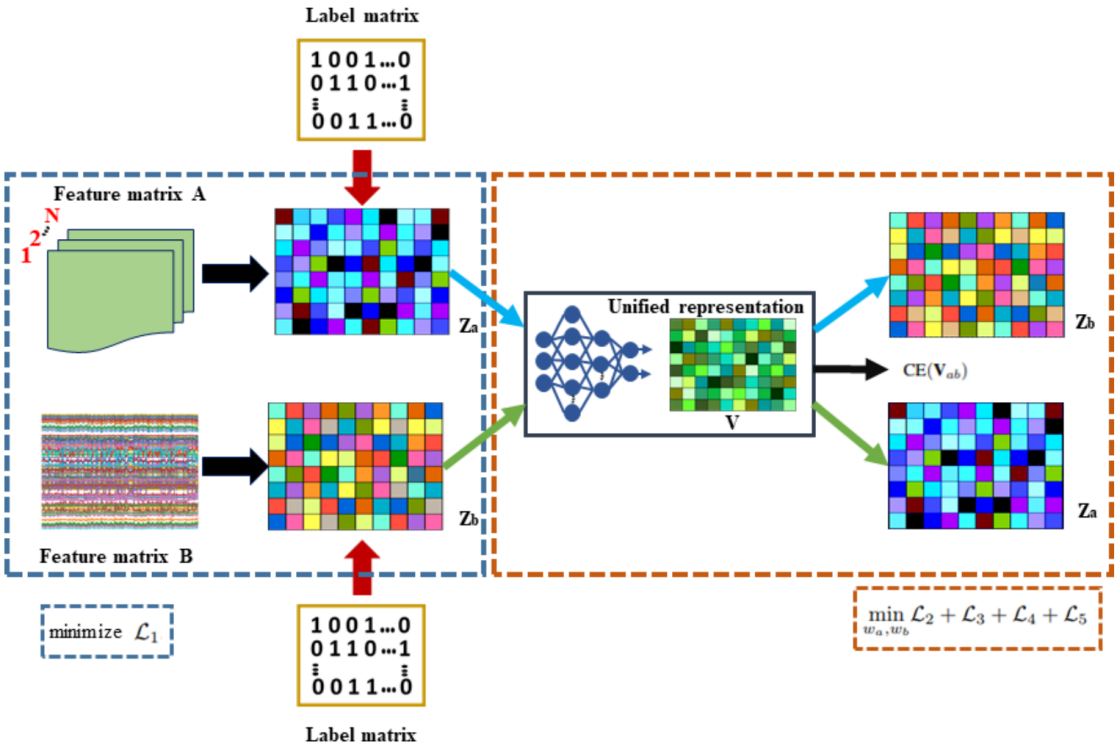


Figure 3.1: CMIR-Net architecture [4]

sisted of a novel cross-modal hashing network and an objective function based on explicitly preserving semantics.

Zhang *et al.* [40] exploited the low-level image features such as textures and spectral features and its content for hyperspectral image retrieval. Chaudhuri *et al.* [4] proposed CMIR-Net network which learns a unified representation from different RS data modalities, and thus improves the cross-modal retrieval performance. The network consists of an encoder-decoder based neural network architecture and is optimized using four loss functions. The architecture is described in Figure 3.1

3.2 Audio-Image Cross-modal RSIR

CMIR-Net proposed by Chaudhuri *et al.* [4] showed that it also works for the audio-image cross-domain retrieval by constructing speech signals manually for land-cover data and testing on it. Chen and Lu [7] used a triplet-based deep hashing network and performed audo-image cross-modal RS image retrieval.

Mao *et al.* [22] curated a large scale RS image-audio dataset containing manually labelled speech captions and performed cross-modal retrieval on it. The authors proposed a deep visual-audio network (DVAN) model to learn a direct

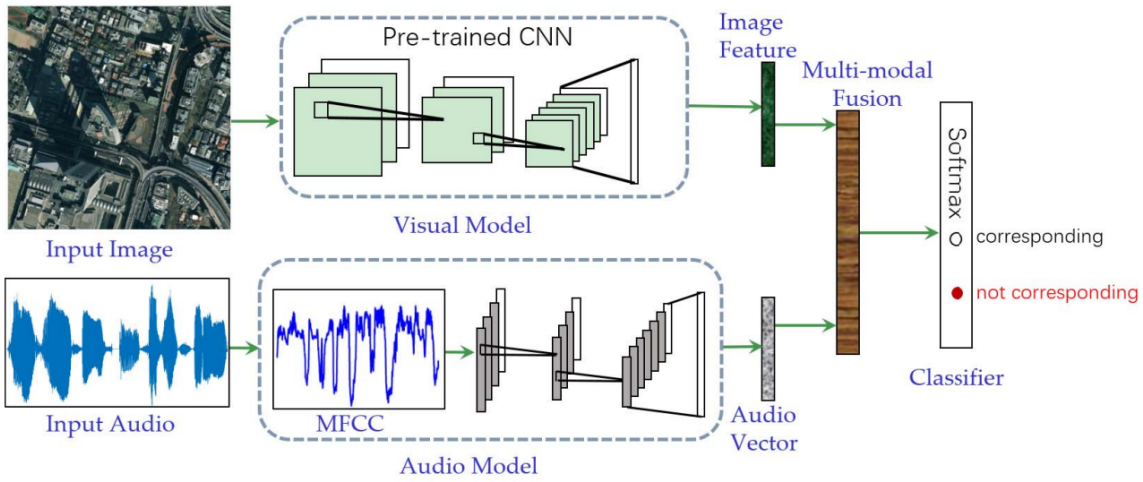


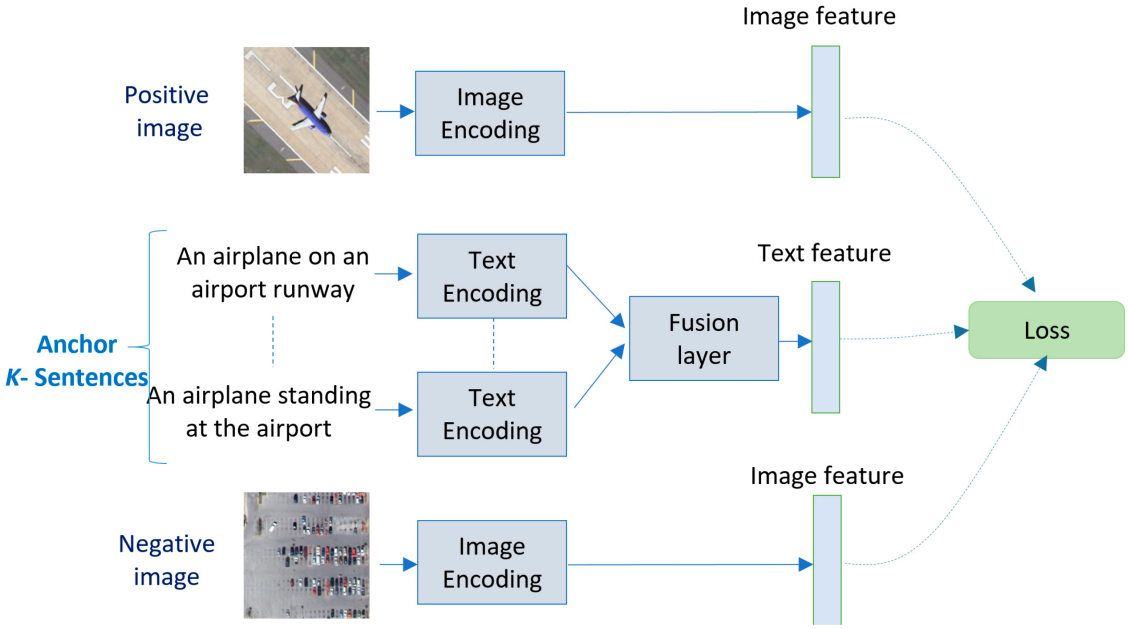
Figure 3.2: DVAN model architecture [22]

correspondence between RS image and its audio caption. Figure 3.2 depicts the DVAN architecture where the image and audio features are passed through a multi-modal fusion layer to obtain a fused embedding to decide if the image-audio pair is related.

3.3 Text-Image Cross-modal RSIR

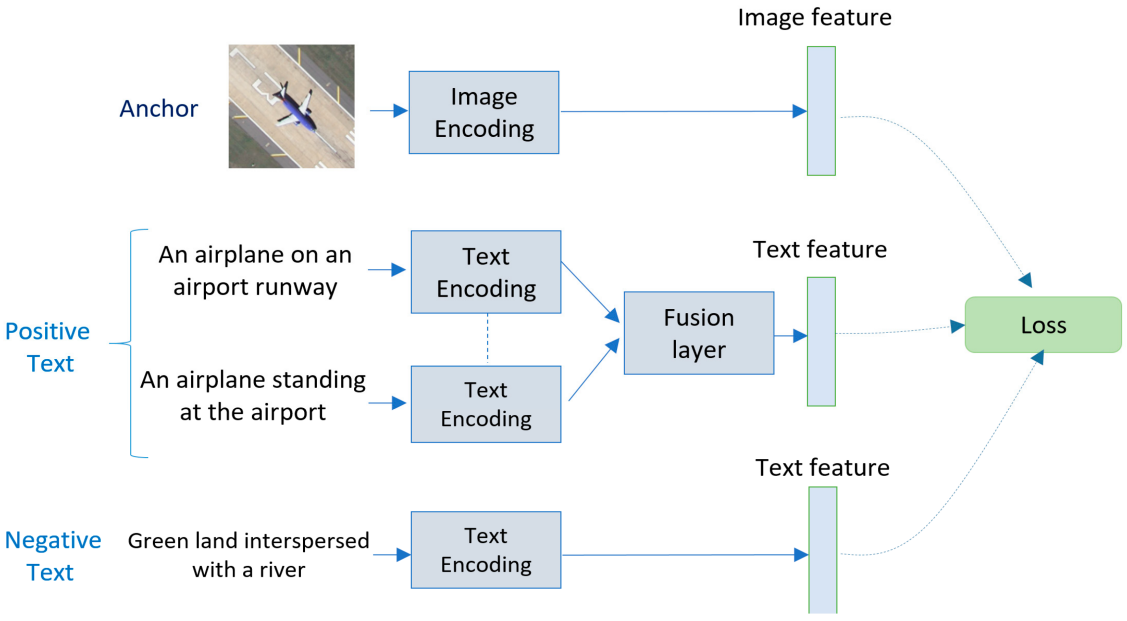
With advancements in automatic image captioning, caption-based image retrieval methods have become quite well-known in the RS domain. Hoxha *et al.* [17] generated and exploited textual descriptions of RS images to learn to describe relationships between object and its attributes present in the images with captions.

Abdullah *et al.* [1] brought to table deep bidirectional triplet network (DBTN) for text-image remote sensing image retrieval, and also constructed a new dataset called TextRS. DBTN was trained using two triplet loss functions, one by keeping the RS image as anchor and the textual descriptions as positive and negative samples, and another by keeping the text as anchor and RS images as positive and negative samples. The flowchart for both the approaches of DBTN are shown in Figure 3.3



(a)

(a) Text as anchor



(b)

(b) Image as anchor

Figure 3.3: Deep Bidirectional Triplet Network architecture

CHAPTER 4
Few-shot Learning

Deep learning models make heavy reliance on labeled data during training. This is not suited to directly learn from a dataset where labeled examples are limited. Thus, a new type of machine learning problem called **few-shot learning** came into existence. The main aim of few-shot learning is to enable a model to perform under practical scenarios where the dataset contains only a limited number of examples with labeled information for the specified machine learning or deep learning task or where data annotation is infeasible [36]. This exactly describes problem setting in this thesis, as we have insufficient training data, and hence we incorporate few-shot learning into our proposed framework.

Specifically, the few-shot learning problem can be termed as a **N-way-K-shot FSL** problem where the data contains only few examples, K, from each of N classes. Existing FSL problems are mainly supervised. Example problems in the computer vision field include image classification [13, 37, 34, 28, 32] and image retrieval [5, 42]. In generalized FSL, the goal is to learn a classifier f where the image/feature x to be recognized at test time may belong to base classes or few-shot classes. Here base class means the classes for which sufficient samples are available and few-shot classes are those which do not have enough samples.

Few-shot learning can be addressed by three perspectives (Figure 4.1) using prior knowledge [36]:

- **Data:** These methods use previous or prior knowledge to augment the few-shot data and increase the number of samples, therefore reducing it to a basic supervised problem. Xian *et al.* [37] proposed a conditional feature generating adversarial network (f-CLSWGAN) for generating CNN features for unseen classes.

- **Algorithm:** These methods use prior knowledge to learn the parameters θ, which gives the best possible hypothesis in the hypothesis space. Prior knowledge provides a good initialization to the model by altering the search

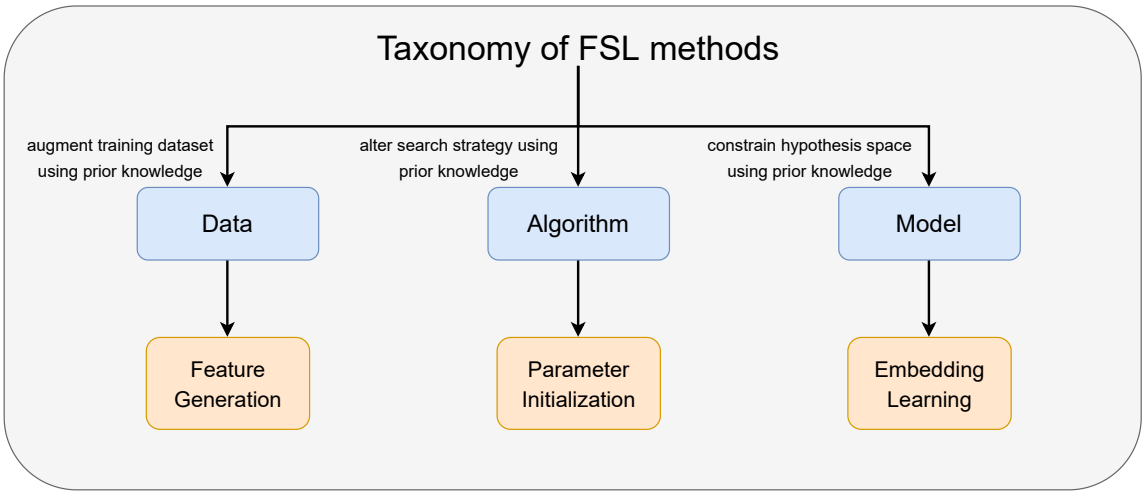


Figure 4.1: A Taxonomy of few-shot learning methods

strategy, or guiding it while searching. Model agnostic meta learning (MAML) by Finn *et al.* [13] meta-learns the parameter set θ, which is then adjusted to obtain a better parameter set ϕ for some task via a few effective gradient descent steps:

$$\phi = \theta - \alpha \nabla_\theta L^S_{train}(\theta) \tag{4.1}$$

Here, $L^S_{train}(\theta)$ is the sum of losses over the training samples in dataset, and α is the step-size.

- **Model:** These methods use prior knowledge to constrain the complexity of hypothesis space, thus making the hypothesis space smaller and resulting in faster convergence from few samples. A method in this domain is called embedding learning which embeds the samples into a lower-dimensional space, such that similar samples are closer and dissimilar can be easily discriminated. Hence, the hypothesis space becomes small, thus requiring fewer training samples.

A taxonomy of the FSL methods is shown in Figure 4.1. The taxonomy describes the types of training strategies for the three ways to approach an FSL problem. Our proposed framework uses a few-shot framework coming under embedding-based learning. Thus, we discuss some embedding-based few-shot learning methods in Section 4.1.

4.1 Embedding/Metric-based Few-Shot Learning

Embedding learning is based on the assumption that if a model or function can determine similarity between two images, it can classify an unseen input in relation to labeled instances seen during the learning process. Embedding learning has three primary components: (i) a function or a model f, which transforms test samples x_{test} to the lower-dimensional embedding space, (ii) a function or a model g, which transforms training samples x_{train} to the embedding space, and (iii) a similarity function s, measuring similarity between lower-dimensional features of x_{test} and x_{train} in the embedding space. The test sample is then assigned to the class of x_{train} with the highest similarity. A common embedding function can also be used for both training and testing. Embedding models are trained by a process called *meta-learning*. Some embedding models are discussed next.

4.1.1 Matching Network

Matching Network [34] trains different embedding functions (f and g) by meta-learning for x_{test} and x_{train}, computes similarity using cosine similarity function s and uses softmax on similarities as the attention mechanism a. f and g can be CNNs or bidirectional LSTMs with different learnable parameter sets depending on input type. Here x_{train} can be called as *support set* and x_{test} can be called as *query set* in the context of retrieval. As shown in Figure 4.2, the test or query image is passed through f to extract deep features from it. Meanwhile the support set is mapped to the function g for extracting their features. Both share the same parameters or weights θ. Thereafter, similarity of query image feature with support set image features are computed. The softmax attention a is applied onto the similarity scores to give more importance to the highest similarity value following which the query image is classified into the category giving the highest similarity.

4.1.2 Prototypical Network (ProtoNet)

Instead of comparing features of x_{test} or query set computed by f with each features of x_{train} or support set computed by g, ProtoNet [28] compares the former with the class prototypes in training set D_{train}. Class prototype for a class n is given by

$$c_n = \frac{1}{K} \sum_{i=1}^{K} g(x_i) \tag{4.2}$$

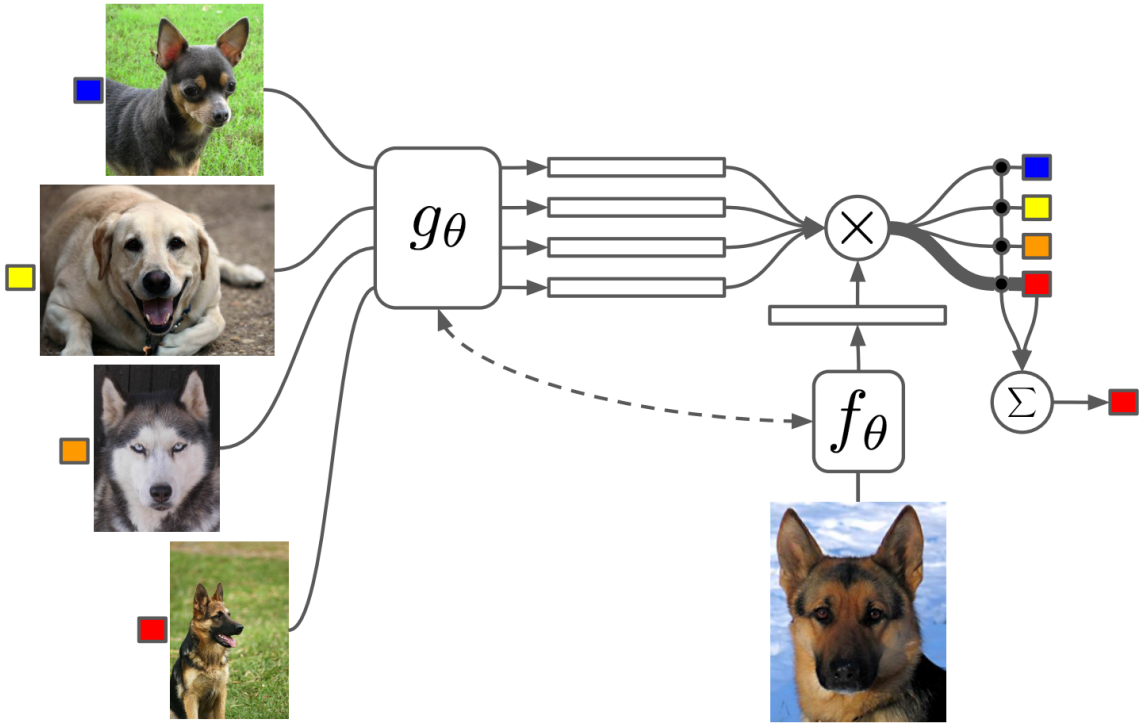


Figure 4.2: Matching Network architecture [34]

Here the functions f and g both have a shared 4-block CNN structure where each block consists of a convolutional layer, a batch normalization layer, a max-pooling layer and a ReLU activation function. ProtoNet uses squared Euclidean distance as its similarity metric. Prototypical network is relatively simple and easy to implement than the matching network. The architecture for prototypical network is shown in Figure 4.3.

4.1.3 Relation Network (RelationNet)

The Relation network [32] also uses a CNN to embed x_{test} and x_{train} to embedding space, then concatenates x_{test} embedding with the x_{train} embeddings. This is then fed to another CNN (non-linear similarity function) to output a similarity score. Its architecture can be seen in Figure 4.4

Some remote sensing image retrieval approaches using few-shot learning include Chaudhuri et al.[5] in which they propose a zero-shot inter-modal retrieval scheme for sketch-based RS retrieval. They even make their bi-modal dataset called earth on canvas (EoC) with original sketches and high resolution RS images. Zero-shot learning aims to solve a specified task without receiving any example during the training phase. This makes the network capable of handling an *unseen* class sample during the inference or testing phase. In [5], the sketch is

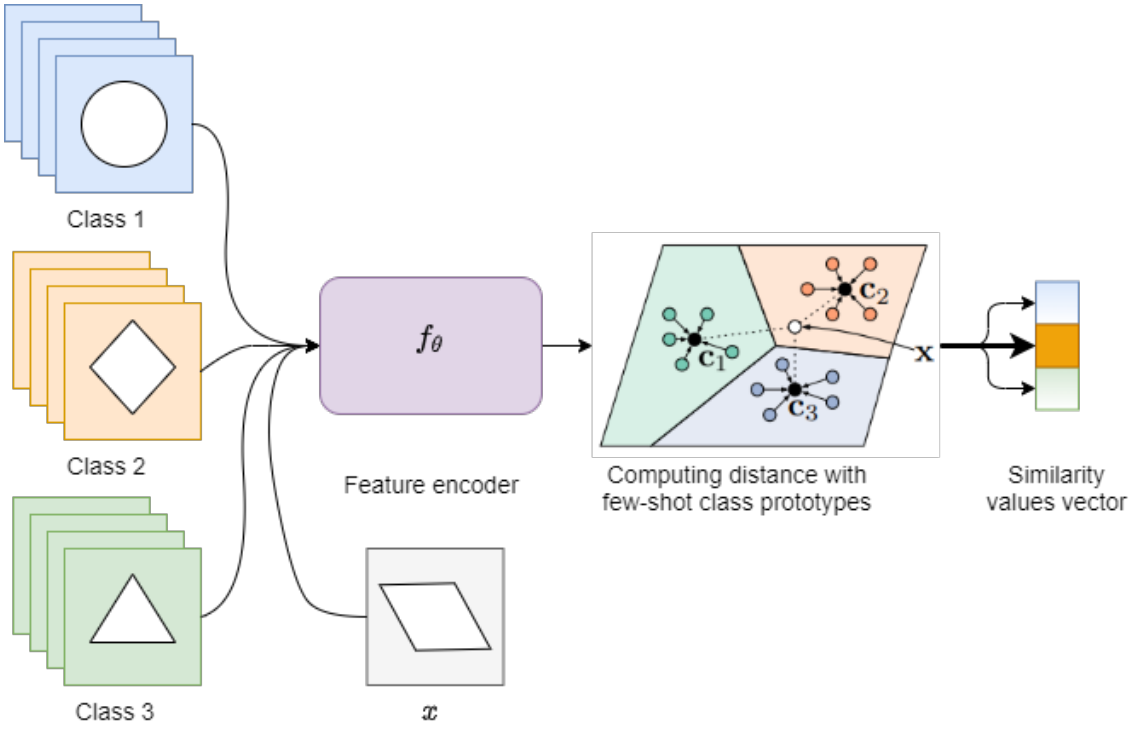


Figure 4.3: Architecture of prototypical network

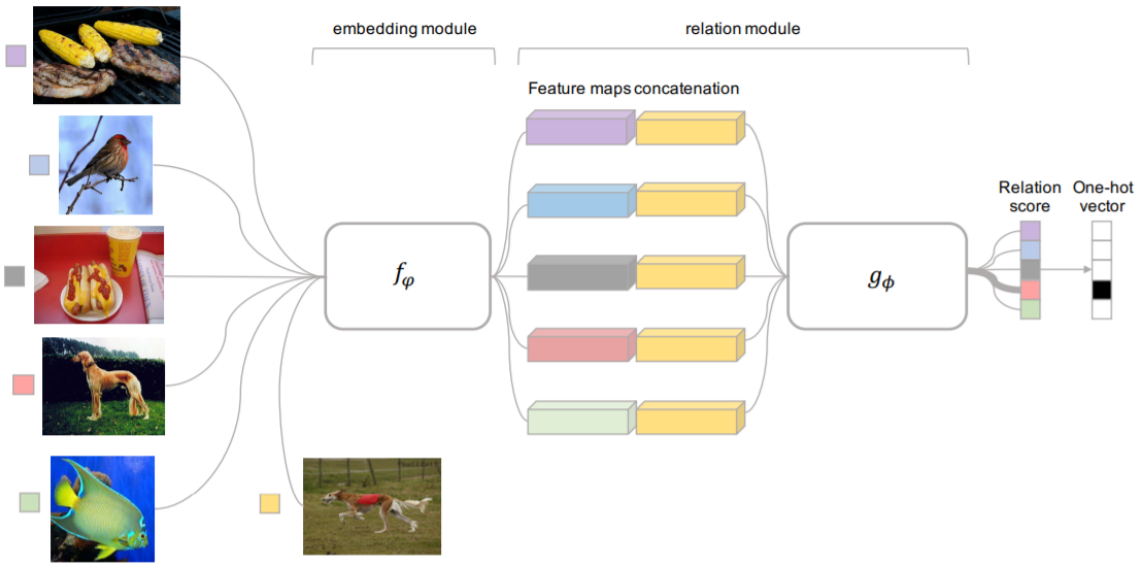


Figure 4.4: Relation network architecture [32]

used to represent the absent training samples. Zhong *et al.* [42] used the MAML few-shot algorithm discussed earlier for retrieving remote sensing images.

CHAPTER 5
Domain Adaptation

Domain adaptation (DA) in deep learning can be considered as a special case of transfer learning (TL) that leverages labeled data in one or more source domains to execute other tasks in a target domain. Domain adaptation techniques came into existence as we often require a machine learning model to be trained for a specific task using training data sampled from one distribution, and deploy on test data sampled from another distribution. Here the training data is from the source domain and the test data is from the target domain. As both the data distributions are different, we see degraded performance of the model on test data even though it performed well on source domain data.

Abundant amount of shallow DA methods have been proposed to solve for the data shift between source and target domains [9, 10, 12, 29]. Advances have also been made in the deep learning field for adaptation. For example, DLID network by Chopra *et al.* [8] trains shared CNN networks simultaneously with a couple of adaptation layers on source and target domains. Their basic idea is to create an interpolating path between source and target domains by creating interpolating domains using CNNs.

Deep adaptation network (DAN) proposed by Long *et al.* [20] tries to minimize the maximum mean discrepancy (MMD) by using multiple kernels applied to the last several fully-connected layers of the convolutional neural network. An overview of DAN is shown in Figure 5.1.

Ganin *et al.* [14] took a standard CNN classification network and attached a deep domain classifier network for domain adaptation with the feature extraction layers. The domain classifier network is attached via a gradient reversal layer which multiplies the gradients of the network by a fixed negative constant during the backpropagation phase. The domain classifier network is trained by learning to classify the feature into its appropriate domain. A negative constant is multiplied to the gradients so that the features generated by the feature extractor are such that they maximize the loss function of the domain classifier network. The

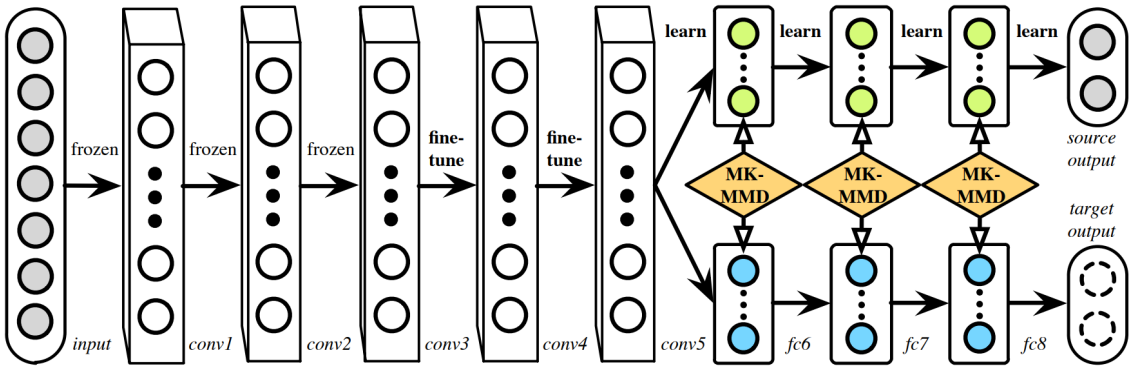


Figure 5.1: An overview of deep adaptation network [20]

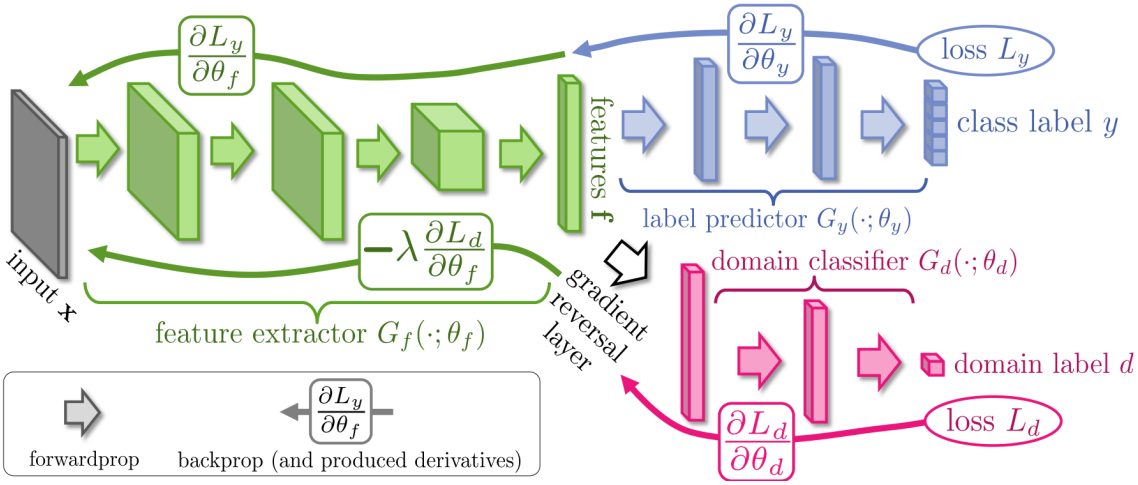


Figure 5.2: Unsupervised deep domain adaptation architecture proposed by Ganin *et al.* [14]

architecture of the network is shown in Figure 5.2

The CORAL algorithm by Sun *et al.* [29] minimizes the distance between covariances C_{source} and C_{target} of source and target domain data respectively by applying a linear transformation A on the source domain covariance matrix:

$$\min_A ||\hat{C}_{source} - C_{target}||_F^2 \tag{5.1}$$

$$= \min_A ||A^T C_{source} A - C_{target}||_F^2 \tag{5.2}$$

where $\hat{C}_{source}$ is the covariance matrix of source features after applying transformation A and $||\cdot||_F^2$ denotes the squared matrix Frobenius norm.

The linear transformation used in CORAL algorithm first whitens the source features using source domain covariance matrix and then re-colors it with the target domain covariance matrix. This is illustrated in Algorithm 1.

Sun and Saenko modified the CORAL algorithm by adapting it for deep learn-

Algorithm 1 CORAL Algorithm [29]

Input: Source Data F_S, Target Data F_T
Output: Transformed Source Data F_S^*

$Cov_S = cov(F_S) + eye(size(F_S, 2))$
$Cov_T = cov(F_T) + eye(size(F_T, 2))$
$F_S = F_S * Cov_S^{\frac{-1}{2}}$ ▷ whitens source
$F_S^* = F_S * Cov_T^{\frac{1}{2}}$ ▷ re-colors with target covariance

ing networks and thus introducing a new DeepCORAL loss [30] for domain adaptation on training and testing data. The DeepCORAL loss also minimizes the distance between covariance matrices of source and target deep features. A detailed explanation of it is discussed in Chapter 6 where we incorporate it in our framework.

CHAPTER 6
Proposed Cross-modal Few-shot Training

To define our architecture in detail, we first introduce some necessary notations.

6.1 Notations

Let A denote the image space of the modality with sufficient/abundant data, S the image space of insufficient/scarce data modality, and $L = \{1, 2, ..., C\}$ be the label space. Furthermore, let a_i be the i-th input image from A, s_i be i-th input image from S, and $y_i \in L$ be the common class label. In a few-shot training episode, there is a small support set of N labeled samples $Sp = \{(a_1, y_1), ..., (a_N, y_N)\}$ and an even smaller set of Q query samples $Sq = \{(s_1, y_1), ..., (s_Q, y_Q)\}$. In our architecture, the few-shot learning scenario is cross-modal while training but single-modal in testing.

6.2 Cross-modal Few-shot Training

An image retrieval framework requires a good feature extractor which can be a part of a classification neural network. Thus, we train our feature extractor for the classification task and use it for retrieval. The architecture of the proposed framework is described in Figure 6.1.

As we are focused on scarce data, we propose to use few-shot learning which is well suited for learning from limited labels. We use a few-shot architecture similar to Prototypical network [28] for training but with ResNet-18 as the feature extractor to gain discriminative representations. The cross-modal meta-training is done using both abundant A and scarce S data modalities but the meta-testing is done only using scarce S data modality since our primary objective is to learn a good feature extractor for scarce data. The prototypes or the mean class vector of

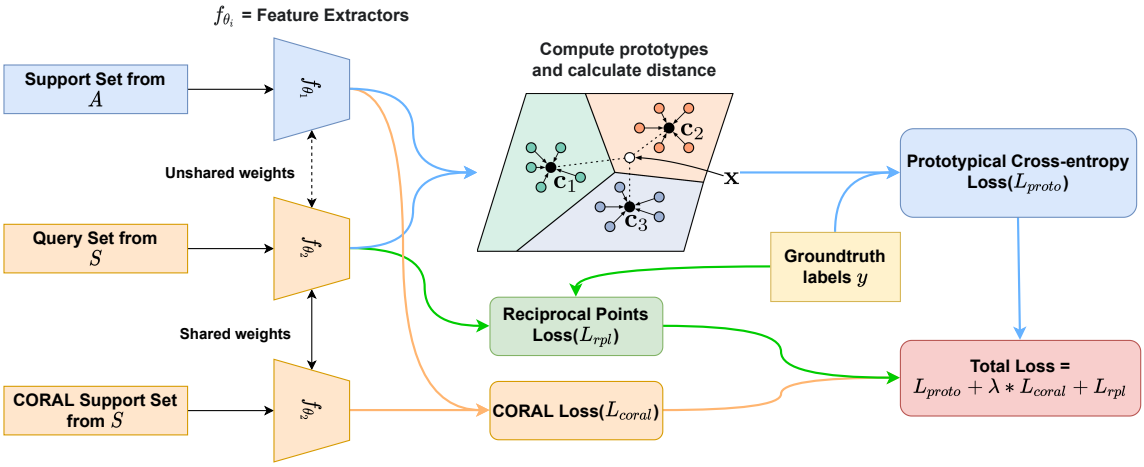


Figure 6.1: Proposed cross-modal few-shot feature extractor training framework

the support set embeddings belonging to that class are computed as follows:

$$l_k = \frac{1}{|Sp_k|} \sum_{(a_i, y_i) \in Sp_k} f_{\theta_1}(a_i) \tag{6.1}$$

where Sp_k suggests set of samples with class label k and f_{θ_1} is the deep feature extractor with θ_1 as the parameters used with support set. The feature vectors of samples from query set are then projected onto the embedding space and distances are computed between every class prototype l_k with the query feature using Euclidean distance function d. Finally the logits are obtained by using negative log softmax of the distances and cross-entropy loss is computed. Thus, we can describe prototypical loss by the below equation:

$$p(y = k|s_i) = \frac{\exp\left(-d(f_{\theta_2}(s_i), l_k)\right)}{\sum_{i=1}^{Q} \exp\left(-d(f_{\theta_2}(s_i), l_k)\right)} \tag{6.2}$$

$$L_{proto} = -\log p(y = k|s_1, s_2, \ldots, s_Q) \tag{6.3}$$

where f_{θ_2} is the second deep feature extractor similar to f_{θ_1} but with θ_2 as its parameters and which is used with the query and CORAL support sets.

Since A and S will be from different modalities, we further propose to use DeepCORAL loss [30] for domain adaptation which is simple to integrate and has achieved better results for the task. DeepCORAL loss minimizes the difference between second order statistics of the source and target domains so that they are aligned well. To integrate it into our few-shot episodic training architecture, we create another support set from S called CORAL support set. The DeepCORAL loss will then be computed between the feature vectors of support set which will

be from A and the CORAL support set from S as:

$$C_A = \frac{1}{n_A - 1}(D_A^T D_A - \frac{1}{n_A}(1^T D_A)^T (1^T D_A)) \tag{6.4}$$

$$C_S = \frac{1}{n_S - 1}(D_S^T D_S - \frac{1}{n_S}(1^T D_S)^T (1^T D_S)) \tag{6.5}$$

$$L_{coral} = \frac{1}{4d^2}||C_A - C_S||_F^2 \tag{6.6}$$

where d is the dimension of feature vector, D_A generated from f_{θ_1} and D_S generated from f_{θ_2} are the features of support set and CORAL support set respectively, and C_A and C_S are covariance matrices computed from them. n_A and n_S are the number of support set and CORAL support set samples in a batch respectively. Thus, we are computing two losses at the same time namely, prototypical cross-entropy loss and DeepCORAL loss. The influence of this loss is controlled by a weight parameter λ. Thus, the combined loss function now can be written as:

$$L_t = L_{proto} + \lambda L_{coral} \tag{6.7}$$

We also incorporate reciprocal points loss [6] into our training. Though its main use is for identifying unknown classes, but it also ensures that known classes are pushed to boundaries and clustered according to their classes which makes the model achieve better performance. Reciprocal points loss is computed by calculating a custom softmax function based on the distance between query feature vectors and computed reciprocal points. Let the set of reciprocal points for class k be denoted as $\mathcal{P}^k = \{p_i^k | i = 1, ..., M\}$ where M is number of reciprocal points per class and $s_q = \{s_1, s_2, \ldots, s_Q\}$ be the query set samples. The distance between query set features and reciprocal points are calculated as follows:

$$d(f_\theta(s_q), \mathcal{P}^k) = \frac{1}{M}\sum_{i=1}^{M} ||f_\theta(s_q) - \mathcal{P}^k||_2^2 \tag{6.8}$$

The custom softmax probabilities are then calculated as follows:

$$p(y = k | s_i, f_\theta, \mathcal{P}^k) = \frac{\exp\left(\gamma d(f_\theta(s_i), \mathcal{P}^k)\right)}{\sum_{i=1}^{Q} \exp\left(\gamma d(f_\theta(s_i), \mathcal{P}^k)\right)} \tag{6.9}$$

$$L_{rpl} = -\log p(y = k | s_q, f_\theta, \mathcal{P}^k) \tag{6.10}$$

where γ is a hyper-parameter. There is also an additional term called open-space

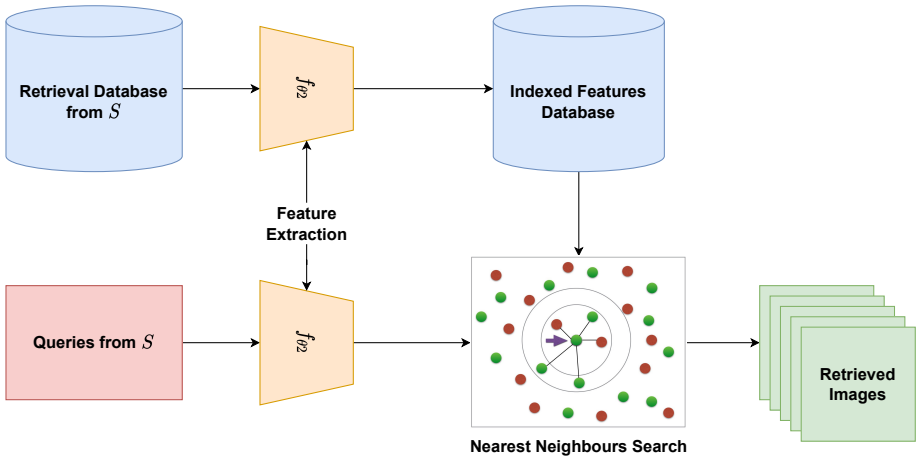


Figure 6.2: Retrieval framework using the scarce modality trained feature extractor and nearest neighbours search

risk added with the loss. However it can be ignored in this setting as the term's purpose is to better distinguish between known and unknown classes. Thus, the final loss for the proposed framework consists of three losses as shown in Equation (6.11).

$$L_t = L_{proto} + \lambda L_{coral} + L_{rpl} \tag{6.11}$$

The architecture of the proposed training framework is shown in Figure 6.1. Here, we train two ResNet-18 feature extractors f_{θ_1} and f_{θ_2} with unshared weights for abundant and scarce modalities respectively. The support set features are extracted from f_{θ_1} model and their prototypes are computed. The query set and CORAL support set features, however, are extracted from f_{θ_2}. The prototypical loss is thus computed using support set prototypes and query set features with true labels, the CORAL loss is computed using support set and CORAL support set features, and the reciprocal points loss is computed using query set features and true labels. All the losses are minimized simultaneously.

After the few-shot training, image retrieval testing is performed on data from S using the trained model f_{θ_2} for extracting features and nearest neighbours search for retrieving images with Euclidean distance as the distance metric which can be written as:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{6.12}$$

where x_i and y_i are feature vector points. The retrieval framework is depicted in Figure 6.2.

CHAPTER 7
Experiments and Results

To validate our cross-modal few-shot retrieval framework, we perform several retrieval experiments using two dual-modality RS image datasets.

7.1 Simulating abundant and scarce data

As per our best knowledge, there is no dataset available which provides us with scarce data in one modality and abundant data in the other. Thus, we simulate data scarcity and abundance from the two dual-modality datasets as per our requirement.

Figure 7.1 provides us a view of how the dataset is partitioned into different splits. First, the original dataset consists of paired remote sensing image patches from two modalities M1 and M2. It is partitioned into a 70:30 split of *abundant-split* and *scarce-split* respectively. Let us consider we choose to simulate modality M2 to be the scarce modality giving us the scarce modality image space S from the scarce-split and modality M1 as the abundant modality giving us the abundant modality image space A from the abundant-split. Thus, query set and CORAL support sets used in the cross-modal few-shot training are created from the simulated scarce modality S and the support set is created from simulated A from the abundant-split. From the abundant-split, we create retrieval testing partition from M2 consisting of a query set consisting of 1000 images distributed equally between the classes and a retrieval set containing the remaining images.

7.2 Datasets

The first dataset we use is CBRSIR-VS [31] consisting of paired very high resolution (VHR) and synthetic aperture radar (SAR) RS images. There are in total 26,901 pairs and 10 class labels namely, industrial buildings, residential buildings,

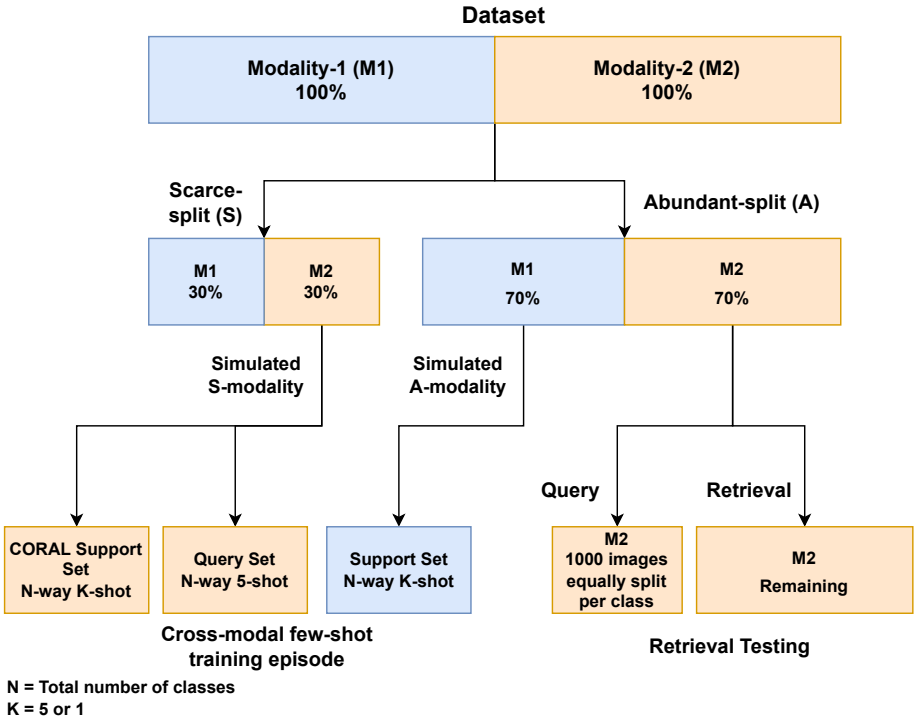


Figure 7.1: Dataset partitioning to simulate scarce and abundant data, using them to create subsets for training and testing

annual crop, permanent crop, river, sea & lake, herbaceous vegetation, highway, pasture, and forest. The geolocation and labels are taken from EuroSAT [16].

Every VHR image is of 256×256 pixels with a spatial resolution of 1 meter and has RGB channel bands, while every SAR image is of 64×64 pixels with a spatial resolution of 10 meters and the spectral band containing only vertically polarized (VV) data. Both the modalities were preprocessed to have a common spatial size of 64×64 pixels with 3 channels. For that purpose, VHR images were resized and SAR images band was triplicated. Data scarcity and abundance for training and testing are simulated as per Figure 7.1. For performing experiments, we once choose VHR modality as scarce and in the other we choose SAR modality as scarce.

The second dataset we perform experiments on is the dual-modality DSRSID [19] dataset. The dataset consists of 80,000 samples of paired panchromatic (PAN) and multispectral (MUL) RS images and 8 land-cover classes namely, aquafarm, cloud, forest, high building, low building, farm land, river, and water.

GF-1 panchromatic and multispectral sensors are used to obtain the images. PAN images are of 256×256 pixels with spatial resolution of 2 meters and single spectral band, while MUL images have a spatial size of 64×64 pixels with a resolution of 8 meters and 4 spectral bands. Also the PAN images are resized to 64×64

pixels and the single band is repeated three more times to match with the number of spectral bands of MUL images. Here also we simulate data scarcity and abundance as per Figure 7.1 and choose PAN modality as scarce in some experiments and MUL modality as scarce in some others.

An overview of the datasets is shown in Table 7.1.

Table 7.1: Overview of Datasets

Dataset	Samples	Classes	Modality-1	Shape	Modality-2	Shape
CBRSIR-VS	26901	10	VHR	(256, 256, 3)	SAR	(64, 64, 1)
DSRSID	80000	8	PAN	(256, 256, 1)	MUL	(64, 64, 4)

7.3 Baselines

To the best of our knowledge, there are no open-source works which performs cross-modal few-shot remote sensing image retrieval. Thus we create two baselines to compare our framework's performance.

- **Uni-modal Baseline:** The first baseline is by training a ResNet18 model for remote sensing image classification using insufficient training data and then using the feature extractor of the model for performing uni-modal RSIR on that modality. As only one-modality is used, it is named *uni-modal baseline*.
- **CORAL Baseline:** Another baseline is created by using a single shared weights ResNet18 model, removing the few-shot learning framework. This training framework will minimize the standard cross-entropy loss and DeepCORAL loss by training a classification network. The model includes cross-modal training by keeping the training set from A as the source domain and the test set from S as the target domain for computing DeepCORAL loss. Evaluation for the CORAL baseline will be done uni-modally as like other setups.

7.4 Implementation details

All modules from cross-modal few-shot training framework to retrieval framework is implemented using PyTorch framework. SGD optimizer was used with a learning rate of 1e-3, 0.9 momentum, and 1e-4 weight decay. The λ value for the DeepCORAL loss was set at 0.3. This value was chosen after some trial and error as we observed that giving it more weight will produce degenerate feature which increases the cross-entropy loss. The weights of prototypical cross-entropy loss

and reciprocal points loss were kept as 1. A learning rate scheduler was used that halved the learning rate after every 20 epochs.

The few-shot training was done for a total of 100 epochs consisting of 100 meta-train episodes. The cross-modal few-shot training episodes were done in N-way-5-shot and N-way-1-shot settings where N is the number of classes in the dataset. Meanwhile few-shot validation episodes were performed in the 5-way-5-shot and 5-way-1-shot settings. During the validation episodes, f_{θ_2} model is used and the CORAL loss and reciprocal points loss are not computed as both the support set and query set are from S. Here, the way or N is set as 5 to evaluate if the model can distinguish a query image, given any combination of classes. This is also to evaluate the model performance in generating discriminative features for the scarce modality. Furthermore, data augmentation was performed on the CORAL support set and the few-shot support set to compute an effective CORAL loss by increasing the number of samples. The more the number of samples, better the model will be able to adapt the source domain to target domain.

7.5 Experimental Setup

We followed two experimental setups and tested it on both the datasets and also reversed the chosen simulated scarce and abundant modalities resulting in a total of 8 experiments to perform.

The first and the primary setup is to use abundant data modality in the support set and the scarce data modality in the query set and CORAL support set while training as described in Figure 6.1.

The second setup is to use abundant data modality images in both the support set and query set while restricting the scarce data modality to only CORAL support set for training. Experiments with this setup are performed to test the hypothesis that without cross-modality the model will not be able to adjust with the scarce modality and in consequence will yield worse results.

During retrieval testing, query and retrieval sets as per Figure 7.1 are used. For each query image from the query set, 100 images are retrieved from the retrieval set and are used for evaluation. In the next section, we discuss the retrieval results of these experiments and compare with the baselines discussed earlier.

7.6 Results

Table 7.2 describes the different types of experimental settings performed using CBRSIR-VS and DSRSID datasets. In the first row, in the training column, the support set is from VHR images, query set and CORAL support set are from SAR images while in the testing column, support set and query set both chosen from SAR images. The train and test set for CORAL Baseline are chosen from VHR images and SAR images respectively and the image retrieval setting for this experiment is to select the query and retrieval sets from SAR images. In this setting, SAR data modality is chosen to be insufficient. Thus, uni-modal baseline in this setting will be trained using SAR data modality.

Table 7.3 is a continuation of Table 7.2 wherein different evaluation metrics are reported with the best metrics highlighted. We chose standard evaluation metrics used for image retrieval namely, mean average precision (mAP), precision at K (P@K). mAP can be calculated from the following equation:

$$mAP = \frac{1}{N} \sum_{k=1}^{N} AP_k \tag{7.1}$$

where AP_k is the average precision of class k and N is the number of classes. Meanwhile, P@K is the proportion of correctly retrieved images out of K total retrieved images.

For both the experimental setups and both 5-shot and 1-shot, we observe that the primary experimental setup where support and query sets are of different modalities gives overall better mAP than the CORAL baseline and uni-modal baseline barring one exception for MUL-MUL setting where P@5 and P@10 values were slightly better. Meanwhile the second setup where the support and query sets are from the same modality failed to beat the baselines. This proves our hypothesis we were testing that the model with the abundant data modality alone is not capable for adapting to the scarce domain based on DeepCORAL and reciprocal points losses. It needs cross-modality or images from the scarce data domain in the meta-training episodes to compute a better embedding space where both the domains are as aligned as possible.

The primary setup showed a 45-65% increase in retrieval performance for VHR and SAR images while for PAN and MUL images, it showed a 26-30% increase. We can also observe from the P@K metrics that most of the relevant retrieved images are in the top. We also observe that by changing the modalities, there is not much of a performance difference in both the datasets which shows its robust-

ness. The good results for the proposed approach are mainly due to introducing cross-modality in the few-shot training process, incorporating deepCORAL loss for domain adaptation between the two modalities, and the reciprocal loss for clustering the same class points together further helping in increasing classification accuracy. Also the marginal performance gain achieved moving from one-shot to five-shot is due to the fact that in the support set only four more image samples per class are added. This is not enough to get a significant gain in the performance.

Table 7.2: Experimental setups for CBRSIR-VS and DSRSID dataset

Sr. No.	Dataset	Cross-modal Few-shot Setting					CORAL Baseline Setting		Retrieval Setting
		Training			Testing				
		SS	QS	CSS	SS	QS	Training	Testing	
1	CBRSIR-VS	VHR	SAR	SAR	SAR	SAR	VHR	SAR	SAR-SAR
2		VHR	VHR	SAR					
3		SAR	VHR	VHR	VHR	VHR	SAR	VHR	VHR-VHR
4		SAR	SAR	VHR					
5	DSRSID	PAN	MUL	MUL	MUL	MUL	PAN	MUL	MUL-MUL
6		PAN	PAN	MUL					
7		MUL	PAN	PAN	PAN	PAN	MUL	PAN	PAN-PAN
8		MUL	MUL	PAN					

SS - Support Set, QS - Query Set, CSS - CORAL Support Set

7.7 Ablation Study

The proposed framework consists of various components. To see their effects in retrieval performance, several ablations were performed by trying out combinations by keeping and removing the components. As the different query and support set modalities setup performed well, ablations were conducted in this setup only. The framework adds two new losses namely DeepCORAL loss and reciprocal points loss. The CORAL baseline we set up removes the few-shot component. This can be considered as an ablation and we observed its results earlier. Thus in this section, we retain the few-shot component and turn on and off the DeepCORAL and reciprocal points losses accordingly. Thus, we get three ablations:

- By turning off only reciprocal points loss.
- By turning off only DeepCORAL loss.

Table 7.3: Image Retrieval results for the setups in Table 7.2

Sr. No.	mAP				P@5			
	Uni-modal Baseline	CORAL Baseline	N-way 5-shot	N-way 1-shot	Uni-modal Baseline	CORAL Baseline	N-way 5-shot	N-way 1-shot
1 2	0.5058	0.3749	**0.5462** 0.307	**0.5698** 0.3067	0.4542	0.364	**0.561** 0.2924	**0.5878** 0.2796
3 4	0.4811	0.3752	**0.6188** 0.328	**0.5981** 0.3392	0.4456	0.3678	**0.608** 0.3104	**0.5926** 0.3326
5 6	0.9384	0.746	**0.9408** 0.6269	**0.8812** 0.6526	**0.9758**	0.738	**0.9753** 0.681	**0.895** 0.6708
7 8	0.8157	0.6798	**0.8875** 0.4942	**0.8627** 0.5146	0.843	0.6872	**0.9522** 0.4922	**0.9033** 0.5215
	P@10				**P@50**			
	Uni-modal Baseline	CORAL Baseline	5-way 5-shot	5-way 1-shot	Uni-modal Baseline	CORAL Baseline	5-way 5-shot	5-way 1-shot
1 2	0.4892	0.3617	**0.5488** 0.3012	**0.5974** 0.285	0.5057	0.3717	**0.5521** 0.3024	**0.5762** 0.2959
3 4	0.4326	0.3593	**0.5999** 0.2895	**0.5636** 0.3132	0.4745	0.3691	**0.6161** 0.3208	**0.5983** 0.3366
5 6	**0.9596**	0.7332	**0.954** 0.6333	**0.8718** 0.64	0.9397	0.7333	**0.9461** 0.6236	**0.8789** 0.6409
7 8	0.8224	0.6901	**0.9112** 0.5005	**0.8663** 0.5206	0.7998	0.6614	**0.885** 0.4863	**0.8616** 0.508

*The Sr. No. column corresponds with the same name column in Table 7.2
N = # of total classes

- By turning off both DeepCORAL and reciprocal points losses.

The mAP and P@K metrics for these ablations and their comparison with the proposed framework with both the losses can be seen in Table 7.5.

From both the tables we observe that the mAP has increased by removing the loss components. For VHR and SAR images, we see that though mAP has increased, the P@K values have reduced. This means there are less relevant images in the top while retrieving. For PAN and MUL images, all the ablations perform better than the proposed framework. This shows that the few-shot and cross-modal components are of significant importance. The weights of DeepCORAL loss and reciprocal points loss can be optimized to further increase the performance but with a risk of dropping P@K values.

One more ablation which can be studied is the variation of the parameter λ used for setting the weight of the DeepCORAL loss L_{coral}. We experimented with the N-way 5-shot setting and compared with λ values 1 and 10. Table 7.6 shows the result metrics of the lambda ablations. Though the results show some improvement in some retrieval settings, they only improve by a small amount. Almost all the values do not deviate much from the proposed approach setting.

Table 7.4: Experimental setups for ablations

Sr. No.	**Dataset**	**Cross-modal Few-shot Setting**					**Retrieval Setting**
		Training			**Testing**		
		SS	**QS**	**CSS**	**SS**	**QS**	
1	**CBRSIR-VS (VHR-SAR images)**	VHR	SAR	SAR	SAR	SAR	SAR–>SAR
2		SAR	VHR	VHR	VHR	VHR	VHR–>VHR
3	**DSRSID (PAN-MUL images)**	PAN	MUL	MUL	MUL	MUL	MUL–>MUL
4		MUL	PAN	PAN	PAN	PAN	PAN–>PAN

SS - Support Set, QS - Query Set, CSS - CORAL Support Set

Table 7.5: Retrieval performance of ablation studies

Sr. No.*	Shots (N-way)	mAP				P@5			
		Proposed	**W/o RP**	**W/o CRL**	**W/o RP-CRL**	**Proposed**	**W/o RP**	**W/o CRL**	**W/o RP-CRL**
1	**5-shot**	0.5462	0.5586	0.5618	**0.5736**	**0.561**	0.4988	0.5143	0.5255
	1-shot	0.5698	0.5694	0.6035	**0.5913**	**0.5878**	0.5232	0.5635	0.5538
2	**5-shot**	0.6188	0.5938	**0.6206**	0.6063	**0.608**	0.541	0.528	0.551
	1-shot	0.5981	0.5817	**0.5987**	0.5909	**0.5926**	0.498	0.5152	0.4785
3	**5-shot**	0.9408	0.9442	0.9455	**0.9448**	0.9753	**0.9845**	0.9697	0.9618
	1-shot	0.8812	0.9162	0.9199	**0.922**	0.895	0.931	0.9505	**0.9598**
4	**5-shot**	**0.8875**	0.8697	0.8874	0.878	0.9522	0.9245	**0.956**	0.9195
	1-shot	0.8627	0.8763	**0.8773**	0.8675	0.9033	0.912	**0.9212**	0.921

		P@10				P@50			
		Proposed	**W/o RP**	**W/o CRL**	**W/o RP-CRL**	**Proposed**	**W/o RP**	**W/o CRL**	**W/o RP-CRL**
1	**5-shot**	**0.5488**	0.51	0.5194	0.531	**0.5521**	0.5294	0.5246	0.5404
	1-shot	**0.5974**	0.523	0.5672	0.5419	**0.5762**	0.5357	0.5724	0.5594
2	**5-shot**	**0.5999**	0.5385	0.562	0.5526	0.6161	0.5936	**0.6202**	0.6188
	1-shot	**0.5636**	0.5139	0.5281	0.5137	0.5983	0.5804	**0.6039**	0.5905
3	**5-shot**	0.954	**0.957**	0.9511	0.9552	0.9461	0.9488	0.9475	**0.95**
	1-shot	0.8718	**0.9312**	0.9305	0.9294	0.8789	0.9218	0.9204	**0.925**
4	**5-shot**	0.9112	0.9079	**0.9144**	0.8991	0.885	0.8656	**0.8934**	0.8743
	1-shot	0.8663	**0.9012**	0.8969	0.899	0.8616	0.8723	**0.8777**	0.8679

*The Sr. No. column corresponds with the same name column in Table 7.4
N = # of total classes
RP = Reciprocal points loss
CRL = DeepCORAL loss

Table 7.6: Retrieval performance of lambda ablations

Sr. No.	Shots (N-way)	mAP			P@5			P@10			P@50		
		$\lambda = 0.3$ (Ours)	$\lambda = 1$	$\lambda = 10$	$\lambda = 0.3$ (Ours)	$\lambda = 1$	$\lambda = 10$	$\lambda = 0.3$ (Ours)	$\lambda = 1$	$\lambda = 10$	$\lambda = 0.3$ (Ours)	$\lambda = 1$	$\lambda = 10$
1	**5-shot**	0.5462	0.5462	**0.5751**	0.561	0.552	**0.5912**	0.5488	0.5386	**0.5977**	0.5521	0.5514	**0.5818**
2	**5-shot**	**0.6188**	0.6148	0.6123	**0.608**	0.5808	0.5786	**0.5999**	0.5906	0.5717	0.6161	**0.6181**	0.6121
3	**5-shot**	0.9408	0.9337	**0.9422**	**0.9753**	0.9625	0.9662	**0.954**	0.9392	0.9498	**0.9461**	0.9447	0.9453
4	**5-shot**	**0.8875**	0.8824	0.8852	**0.9522**	0.9332	0.942	0.9112	0.9129	**0.9169**	0.885	0.8801	**0.8852**

*The Sr. No. column corresponds with the same name column in Table 7.4

CHAPTER 8
Conclusion and Future Work

In this thesis, we presented a novel cross-modal few-shot remote sensing image retrieval framework for when limited data is available for one modality. This report also validated the framework on two datasets and found them to be achieving better results and beating the baseline suggested. Furthermore, the framework proves to be robust as there is not much performance degradation when the modalities are swapped. The ablation study showed that the few-shot and cross-modal components were the main factors resulting in a good retrieval performance. The DeepCORAL loss and reciprocal points loss though not having significant importance, can improve or degrade the performance depending on the datasets.

For the future work, different few-shot learning networks like MatchingNet, RelationNet, MAML, etc. can be trained and evaluated instead of only Prototypical network. The hyperparameters like λ for the DeepCORAL loss can be varied to see which parameters work the best. As no similar work was found, the framework can be compared with existing uni-modal retrieval methods by evaluating them on the same scarce modality experimental setting.

References

[1] T. Abdullah, Y. Bazi, M. M. Al Rahhal, M. L. Mekhalfi, L. Rangarajan, and M. Zuair. Textrs: Deep bidirectional triplet network for matching text to remote sensing images. *Remote Sensing*, 12(3), 2020.

[2] E. Aptoula. Bag of morphological words for content-based geographical retrieval. In *2014 12th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 1–5, 2014.

[3] P. Bosilj, E. Aptoula, S. Lefèvre, and E. Kijak. Retrieval of remote sensing images with pattern spectra descriptors. *ISPRS International Journal of Geo-Information*, 5(12), 2016.

[4] U. Chaudhuri, B. Banerjee, A. Bhattacharya, and M. Datcu. Cmir-net : A deep learning based model for cross-modal retrieval in remote sensing. *Pattern Recognition Letters*, 131:456–462, 2020.

[5] U. Chaudhuri, B. Banerjee, A. Bhattacharya, and M. Datcu. A zero-shot sketch-based intermodal object retrieval scheme for remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, pages 1–5, 2021.

[6] G. Chen, L. Qiao, Y. Shi, P. Peng, J. Li, T. Huang, S. Pu, and Y. Tian. Learning open set network with discriminative reciprocal points. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III*, page 507–522, Berlin, Heidelberg, 2020. Springer-Verlag.

[7] Y. Chen and X. Lu. A deep hashing technique for remote sensing image-sound retrieval. *Remote Sensing*, 12(1), 2020.

[8] S. Chopra, S. Balakrishnan, and R. Gopalan. Dlid: Deep learning for domain adaptation by interpolating between domains. In *ICML Workshop (2013)*, 2013.

[9] L. Duan, I. Tsang, D. Xu, and T.-S. Chua. Domain adaptation from multiple sources via auxiliary classifiers. In *Proceedings of the 26th International Conference On Machine Learning, ICML 2009*, volume 382, page 37, 01 2009.

[10] L. Duan, I. W. Tsang, and D. Xu. Domain transfer multiple kernel learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):465–479, 2012.

[11] S. R. Dubey. A decade survey of content based image retrieval using deep learning. *IEEE Transactions on Circuits and Systems for Video Technology*, page 1–1, 2021.

[12] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *2013 IEEE International Conference on Computer Vision*, pages 2960–2967, 2013.

[13] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1126–1135, 2017.

[14] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation, 2014.

[15] Y. Gu, Y. Wang, and Y. Li. A survey on deep learning-driven remote sensing image scene understanding: Scene classification, scene retrieval and scene-guided object detection. *Applied Sciences*, 9(10), 2019.

[16] P. Helber, B. Bischke, A. Dengel, and D. Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification, 2017.

[17] G. Hoxha, F. Melgani, and B. Demir. Toward remote sensing image retrieval under a deep image captioning perspective. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:4462–4475, 2020.

[18] H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3304–3311, 2010.

[19] Y. Li, Y. Zhang, X. Huang, and J. Ma. Learning source-invariant deep hashing convolutional neural networks for cross-source remote sensing image retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 56(11):6521–6536, 2018.

[20] M. Long, Y. Cao, J. Wang, and M. I. Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 97–105. JMLR.org, 2015.

[21] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, Nov 2004.

[22] G. Mao, Y. Yuan, and L. Xiaoqiang. Deep cross-modal retrieval for remote sensing image and audio. In *2018 10th IAPR Workshop on Pattern Recognition in Remote Sensing (PRRS)*, pages 1–7, 2018.

[23] P. Napoletano. Visual descriptors for content-based retrieval of remote-sensing images. *International Journal of Remote Sensing*, 39(5):1343–1376, 2018.

[24] D. Peijun, C. Yunhao, T. Hong, and F. Tao. Study on content-based remote sensing image retrieval. In *Proceedings. 2005 IEEE International Geoscience and Remote Sensing Symposium, 2005. IGARSS '05.*, volume 2, pages 4 pp.–, 2005.

[25] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[26] M. Quartulli and I. G. Olaizola. A review of eo image information mining, 2012.

[27] Sivic and Zisserman. Video google: a text retrieval approach to object matching in videos. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 1470–1477 vol.2, 2003.

[28] J. Snell, K. Swersky, and R. Zemel. Prototypical networks for few-shot learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[29] B. Sun, J. Feng, and K. Saenko. Return of frustratingly easy domain adaptation. In *AAAI*, 2016.

[30] B. Sun and K. Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV 2016 Workshops*, 2016.

[31] Y. Sun, S. Feng, Y. Ye, X. Li, J. Kang, Z. Huang, and C. Luo. Multisensor fusion and explicit semantic preserving-based deep hashing for cross-modal

remote sensing image retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022.

[32] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[33] X.-Y. Tong, G.-S. Xia, F. Hu, Y. Zhong, M. Datcu, and L. Zhang. Exploiting deep features for remote sensing image retrieval: A systematic investigation. *IEEE Transactions on Big Data*, 6(3):507–521, 2020.

[34] O. Vinyals, C. Blundell, T. Lillicrap, k. kavukcuoglu, and D. Wierstra. Matching networks for one shot learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[35] X. Wang, R. B. Girshick, A. Gupta, and K. He. Non-local neural networks. *CoRR*, abs/1711.07971, 2017.

[36] Y. Wang, Q. Yao, J. Kwok, and L. M. Ni. Generalizing from a few examples: A survey on few-shot learning, 2020.

[37] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata. Feature generating networks for zero-shot learning, 2018.

[38] Y. Yang and S. Newsam. Geographic image retrieval using local invariant features. *IEEE Transactions on Geoscience and Remote Sensing*, 51(2):818–832, 2013.

[39] F. Ye, H. Xiao, X. Zhao, M. Dong, W. Luo, and W. Min. Remote sensing image retrieval using convolutional neural network features and weighted distance. *IEEE Geoscience and Remote Sensing Letters*, 15(10):1535–1539, 2018.

[40] J. Zhang, W. Geng, X. Liang, J. Li, L. Zhuo, and Q. Zhou. Hyperspectral remote sensing image retrieval system using spectral and texture features. *Applied Optics*, 56(16):4785, May 2017.

[41] 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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:2711–2723, 2021.

[42] Q. Zhong, L. Chen, and Y. Qian. Few-shot learning for remote sensing image retrieval with maml. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 2446–2450, 2020.

[43] W. Zhou, S. Newsam, C. Li, and Z. Shao. Learning low dimensional convolutional neural networks for high-resolution remote sensing image retrieval. *Remote Sensing*, 9(5), 2017.

[44] Z. Zhuo and Z. Zhou. Remote sensing image retrieval with gabor-ca-resnet and split-based deep feature transform network. *Remote Sensing*, 13(5), 2021.