

https://dspace.org/

Cross-domain few-shot classification for remote sensing imagery

Pillay, Christopher Wayne

2025-04
University of the Witwatersrand, Johannesburg
https://hdl.handle.net/10539/47798

Pillay, Christopher Wayne. (2025). Cross-domain few-shot classification for remote sensing imagery. [Master's dissertation, University of the Witwatersrand, Johannesburg]. WIReDSpace. https://hdl.handle.net/10539/47798

Downloaded from a DSpace repository. https://dspace.org/

Cross-domain few-shot classification for remote sensing imagery

Masters Research Dissertation

School of Computer Science & Applied Mathematics
University of the Witwatersrand

Christopher Wayne Pillay
1362077

Supervised by Dr Hairong Bau

April 11, 2025



A thesis submitted to the Faculty of Science, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Bachelor of Science with Masters

Abstract

Deep learning has proven highly effective for scene classification tasks when substantial quantities of labelled data are accessible. However, performance decreases when applied to domains such as remote sensing which typically possess a limited quantity of labelled data across available datasets. Few-shot learning has been developed as one of the promising solutions to this problem. It has the ability to recognise new categories with minimal labelled examples, but it assumes that the training and testing data will exhibit identical feature distributions. This assumption is unrealistic in real-world contexts where data can originate from different domains and poses a challenge when a significant domain shift exists between the training and testing data. This dissertation aims to address these limitations by proposing the Cross-Domain Attention Network (CDAN). It is a network designed specifically to solve the issues that arise when there is a limited quantity of labelled data available and a significant domain shift exists between the training and testing data. The network proposed consists of a prototypical network as the base and three additions that contribute to the accurate scene classification of remote sensing imagery. Firstly, a cross-domain data augmentation technique is proposed with few-shot learning to reduce domain shift. The cross-domain data augmentation technique facilitates enhanced knowledge transfer between domains and increases the adaptation ability of the network, whereas few-shot learning reduces the network's reliance on large labelled datasets. Secondly, a dynamic and focused attention module is proposed to improve discriminative capacity of the network by increasing the focus on important channels and spatial regions within images during training. Thirdly, an adaptive task aware loss is proposed to further enhance the network's adaptive capacity by leveraging information in few-shot training tasks. Extensive experiments are carried out with different remote imaging classification datasets (RSI-CB, AID and NWPU-RESISC45) to prove that the proposed network alleviates concerns in a cross-domain few-shot (CDFS) classification setting.

Acknowledgements

Computations were generated by utilising the High Performance Computing facility supplied by the Mathematical Sciences Support department at the University of the Witwatersrand.

Contents

Preface

Abstract i

Acknowledgements ii

Table of Contents iii

List of Figures vii

List of Tables x

1 Introduction **1**

1.1 Background 1

1.2 Problem Statement 3

1.3 Research Aim and Objectives 4

1.3.1 Research Aim 4

1.3.2 Research Objectives 4

1.4 Research Methodology 5

1.5 Main Findings 5

1.6 Research Limitations and Assumptions 6

1.7 Summary 6

2 Literature Review **8**

2.1 Introduction 8

2.2 Remote Sensing Imagery 9
2.3 Convolutional Neural Networks 10
2.3.1 Nature of CNNs 10
2.3.2 Attention Mechanisms 11
2.3.3 Metric Learning 13
2.3.4 Loss Functions 14
2.4 Learning Paradigms 15
2.4.1 Preliminary 15
2.4.2 Supervised Learning 17
2.4.3 Unsupervised Learning 18
2.4.4 Generative Learning 19
2.4.5 Few-Shot Learning 20
2.5 Data Augmentation 25
2.5.1 Cutout 26
2.5.2 Random Erasing 26
2.5.3 Mixup 27
2.5.4 CutMix 27
2.6 Summary 29

3 Methodology **30**

3.1 Introduction 30
3.2 Research Hypothesis 30
3.3 Data 31
3.4 Data Description 31
3.4.1 Aerial Image Dataset (AID) 31
3.4.2 Remote Sensing Image Scene Classification dataset by Northwestern Polytechnical University (NWPU-RESISC45) 32

3.4.3 Remote Sensing Image Classification Benchmark (RSI-CB) 32
3.4.4 Mini-ImageNet 32
3.5 Data Pre-processing 32
3.6 Proposed Network 33
3.6.1 Network Outline 33
3.6.2 Cross-Domain Region MixUp 33
3.6.3 Dynamically Focused Attention Module 36
3.6.4 Adaptive Loss Function 41
3.7 Experimental Setup 44
3.7.1 Baseline Network and Comparison Architectures 45
3.7.2 Implementation Details 45
3.7.3 Evaluation Details 46
3.8 Research Limitations and Assumptions 47
3.9 Summary 48

4 Results and Discussion **49**
4.1 Introduction 49
4.2 AID 49
4.2.1 Results and Discussion 49
4.3 NWPU-RESISC45 53
4.3.1 Results and Discussion 53
4.4 RSI-CB 56
4.4.1 Results and Discussion 56
4.5 Effect of N-way Variations 60
4.6 Effect of Scaling Factor α_2 62
4.7 Effect of Network Depth 63

4.8 Ablation Study 66
4.8.1 Effect of Cross-Domain Region MixUp 66
4.8.2 Effect of Dynamically Focused Attention Module 67
4.8.3 Effect of all Additions 70
4.9 Summary 71

5 Conclusion **73**
5.1 Research Summary 73
5.2 Limitations of the Study 75
5.3 Future Work 75

A Additional Information **76**
A.1 Source Dataset Split 76
A.2 Target Dataset Splits 77
A.3 ResNet18 Structure 78
A.4 Architecture Parameters 79

References **88**

List of Figures

2.1 Samples from multiple classes in the Aerial Image Dataset. 10

2.2 A visualisation of how CBAM is set up between ResBlocks [Woo *et al.* 2018]. 12

2.3 Comparison of dataset use in supervised classification vs few-shot classification. Adapted from Zhang *et al.* [2021]. 21

2.4 Comparison of existing types of CDFS learning approaches, which are split into multiple and single source categories. Adapted from Wang *et al.* [2023a]. 23

2.5 Single source CDFS learning visualisation of the training and testing procedure with a network. Adapted from Sa *et al.* [2022]. 25

2.6 Cutout visualisation with remote sensing imagery. 26

2.7 Random Erasing visualisation with remote sensing imagery. 27

2.8 MixUp visualisation with remote sensing and natural imagery. 28

2.9 CutMix visualisation with remote sensing and natural imagery. 29

3.1 The diagram depicts the structure of the input data and the CNN, which emphasises the training and testing procedure used to solve CDFS classification. 34

3.2 This diagram depicts the process on input images using CDRM. The first row represents the process that the target domain input image undergoes, while the second row represents changes that occur with respect to the source domain input image. 36

3.3 The DFAM contains the dynamic channel attention module and hierarchical spatial attention module. The diagram depicts the process where DFAM takes an input feature and produces the refined feature output. . 37

3.4 The diagram represents the Dynamic Channel Attention Module structure. It shows how the input feature is utilised to generate $\mathbf{M_c}$ in the top branch and $\mathbf{M_d}$ in the lower branch to produce the final output feature map. 39

3.5 The diagram depicts the structure of the Hierarchical Spatial Attention Module. It shows how the channel refined feature is utilised to generate the output feature map $\mathbf{M_{spatial}}$. 40

4.1 CDFS classification F1-score test results on different architectures for AID with 5-Way K-Shot settings, where K={1, 5, 10} 52

4.2 CDFS classification F1-score test results on different architectures for NWPU with 5-Way K-Shot settings, where K={1, 5, 10} 55

4.3 CDFS classification F1-score test results on different architectures for RSI-CB with 5-Way K-Shot settings, where K={1, 5, 10} 58

4.4 CDFS classification accuracy for the Baseline and CDAN on AID with N-way K-shot settings. 61

4.5 CDFS classification accuracy for the Baseline and CDAN on NWPU-RESISC45 with N-way K-shot settings. 61

4.6 CDFS classification accuracy for the Baseline and CDAN on RSI-CB with N-way K-shot settings. 62

4.7 CDFS classification accuracy for all target datasets using CDAN with 5-way 5-shot settings while varying the scaling factor α_2. 63

4.8 CDFS classification network depth test results on different target datasets using CDAN with 5-Way 1-Shot settings. 64

4.9 CDFS classification network depth test results on different target datasets using CDAN with 5-way 5-shot settings. 65

4.10 CDFS classification network depth test results on different target datasets using CDAN with 5-way 10-shot settings. 65

4.11 Diagram of original scene images and attention maps derived from Grad-Cam using ProtoNet and ProtoNet with DFAM architectures. Classes are chosen from the AID novel/test set. 68

4.12 Diagram of original scene images and attention maps derived from Grad-Cam using ProtoNet and ProtoNet with DFAM architectures. Classes are chosen from the NWPU-RESISC45 novel/test set. 69

4.13 Diagram of original scene images and attention maps derived from GradCam using ProtoNet and ProtoNet with DFAM architectures. Classes are chosen from the RSI-CB novel/test set. 69

A.1 The ResNet18 CNN structure. 78

List of Tables

4.1 CDFS classification accuracy test results on different architectures for AID with 5-Way K-Shot settings. 50

4.2 CDFS classification accuracy test results on different architectures for NWPU-RESISC45 with 5-Way K-Shot settings. 53

4.3 CDFS classification accuracy test results on different architectures for RSI-CB with 5-Way K-Shot settings. 56

4.4 Ablation study test results highlighting the effect of CDRM on ProtoNet's classification accuracy in 5-way K-shot tasks for different target datasets. 67

4.5 Ablation study test results highlighting the effect of DFAM on ProtoNet's classification accuracy in 5-way K-shot tasks for different target datasets. 67

4.6 CDFS classification accuracy test results for main ablation study with 5-way K-shot settings using AID. 71

4.7 CDFS classification accuracy test results for main ablation study with 5-way K-shot settings using NWPU-RESISC45. 71

4.8 CDFS classification accuracy test results for main ablation study with 5-way K-shot settings using RSI-CB. 71

A.1 Dataset split for Mini-ImageNet. 76

A.2 Dataset split for AID. 77

A.3 Dataset split for NWPU-RESISC45. 77

A.4 Dataset split for RSI-CB. 77

A.5 Total number of learnable parameters for different architectures. 79

Acronyms

AID Aerial Image Dataset.

ATAL Adaptive Task Aware Loss.

CBAM Convolutional Block Attention Module.

CDAN Cross-Domain Attention Network.

CDFS Cross-Domain Few-Shot.

CDRM Cross-Domain Region MixUp.

CNN Convolutional Neural Network.

CSM Channel Selector Module.

DCAM Dynamic Channel Attention Module.

DFAM Dynamically Focused Attention Module.

HSAM Hierarchical Spatial Attention Module.

NWPU-RESISC45 Remote Sensing Image Scene Classification dataset by Northwestern Polytechnical University.

RSI-CB Remote Sensing Image Classification Benchmark.

Chapter 1

Introduction

This dissertation proposes the Cross-Domain Attention Network (CDAN) to solve few-shot scene classification tasks on remote sensing imagery. CDAN consists of a prototypical network as a base with three complementary methods to aid in the scene classification of remote sensing imagery. The additions involve the creation of a Convolutional Neural Network (CNN) that is optimised to perform well in a few-shot setting where there is significant domain shift. This is achieved by focusing on improving the overall adaptability of the network. This chapter introduces the topic and outlines the structure of the dissertation.

1.1 Background

Scene classification is a process which aims to sort scenes from images or photos according to their respective category [King *et al.* 2016]. This is a major research subject in the computer vision and remote sensing domain since it enables the characterisation of human activity and land cover in different areas [Gong *et al.* 2018]. Other benefits of scene classification on remote sensing imagery include environmental monitoring, urban flood modelling, urban water and gas pollution [Huang *et al.* 2019; Wang *et al.* 2018b; Tane *et al.* 2018].

The scene classification of remote sensing images is particularly challenging due to the inherent characteristics of the images, since the entities being represented in the images vary in scale and orientation [Mekhalfi *et al.* 2015]. For example, an urban area could be represented at different levels of detail, which may or may not lead to different objects being distinguishable. Similarly, lighting or seasonal variations may lead to natural landscapes appearing vastly different depending on the viewing angle. Furthermore, remote sensing image data is complex and possesses large amounts of information, which includes the number of objects a scene contains (dense urban areas compared to rural farmlands), spatial relationships that exist between the objects in a scene and additional high-level information [Li *et al.* 2020b]. These characteristics lead

to various challenges when dealing with remote sensing images in scene classification.

A prominent challenge is high intra-class variance within remote sensing data. For example, this occurs when images of natural landscapes, such as land for agricultural use, will vary throughout the year depending on the season or soil condition, yet these images would be part of one class. Another major challenge is high inter-class similarity within remote sensing data which further complicates classification tasks. For example, this case can occur between urban and industrial areas that share similar features such as roads and building shapes, which would make it challenging for a model to differentiate between the two classes [Li *et al.* 2020b; Guo *et al.* 2019]. Additional challenges related to remote sensing imagery include the resolution of the data, as processing and analysis of these types of large-scale datasets require substantial computational resources [Ma *et al.* 2015].

Deep learning methods, in particular supervised approaches, have been a popular and highly effective approach in solving scene classification of remote sensing imagery as well as other computer vision related problems [Wang and others 2016; Zeng *et al.* 2021; Chaib *et al.* 2017; Gong *et al.* 2018; Li *et al.* 2017a; Cheng *et al.* 2017; Ma *et al.* 2019a; Li *et al.* 2020c]. The first noticeable breakthrough with deep learning to solve computer vision problems was achieved with ImageNet [Krizhevsky *et al.* 2012] which utilised a CNN. CNNs are highly effective as they possess the ability to automatically extract and learn hierarchical features at lower levels (e.g., edges and textures) and higher levels (e.g., scenes and objects) [Gu *et al.* 2018]. This quality allows CNNs to automatically bridge the semantic gap through hierarchical abstraction of images, which is crucial for scene classification tasks [Ma *et al.* 2019a]. These properties enable effective classification of remote sensing imagery.

Despite the major success associated with supervised deep learning approaches in computer vision applications, a substantial challenge persists with a supervised approach. The major issue is that supervised deep learning requires large, labelled datasets for the effective training of CNNs [Ma *et al.* 2019b]. The large amounts of data help models learn intricate patterns and enable them to generalise well to new, unseen examples. In the remote sensing field, the generation of large, labelled datasets remains difficult, since the labelling process is time-consuming and expensive given the intricacy of scenes within the data [Alajaji *et al.* 2020]. Therefore, few-shot learning has been a popular alternative to supervised deep learning. It works to solve the issue of data scarcity through task-based training to mimic how humans learn to recognise new categories with a few examples. However, few-shot learning assumes that the training and testing data originates from the same domain. This assumption can lead to a lack of performance once the trained model is applied to other domains, because of the shift in feature distributions that exists between data from different domains [Chen *et al.* 2019].

The cross-domain few-shot (CDFS) research field seeks to overcome the challenges associated with regular few-shot learning. It still aims to reduce dependencies on large labelled datasets but also focuses on effective knowledge transfer from one domain

to another [Guo *et al.* 2020]. This property allows a model to generalise across different domains without requiring intensive training on new domains. It is achieved by utilising shared representations between domains to guarantee the resilience of the model when it encounters variations in the feature distributions between the domains. Therefore, the models become more adaptable since cross-domain learning can mitigate domain shift that occurs when utilising datasets from disparate domains [Guo *et al.* 2020; Tseng *et al.* 2020; Li *et al.* 2021; Wang *et al.* 2023ba].

The challenges related to remote sensing data, supervised and few-shot learning have resulted in a surge of research focusing on the CDFS learning field [Chen *et al.* 2019; Guo *et al.* 2020; Tseng *et al.* 2020; Li *et al.* 2021; Wang *et al.* 2023b]. This field of research offers solutions to the limitations of the traditional methods. It is an advantageous method for diverse and limited labelled data domains like remote sensing. However, there is a limited amount of research that solves CDFS scene classification problems with remote sensing imagery, which will be the primary focus in this dissertation.

1.2 Problem Statement

Existing scene classification approaches are numerous, with the most robust models utilising large amounts of labelled training data to recognise a class within the dataset. This is particularly applicable for remote sensing imagery, due to complex characteristics of the data. The need for large datasets to perform effective scene classification is a challenge when dealing with remote sensing imagery due to cost and time requirements for the creation of a large, labelled dataset. Additionally, most existing models struggle to recognise novel classes, especially when they originate from a different domain. This also constrains the practical application of models in real-world contexts where domain shifts are common. In summary, these issues may be formulated as a CDFS learning problem on remote sensing imagery, where the goal is effective scene classification.

Therefore, a wider variety of approaches are required to create a robust network to address the problem. This research introduces three complementary methods that are combined with a prototypical network to solve CDFS scene classification on remote sensing datasets. The first method is a cross-domain data augmentation method, Cross-Domain Region MixUp (CDRM), which is used to enhance adaptation ability of a CNN and further decrease any existing dependencies on large datasets. The second method is an attention module, Dynamically Focused Attention Module (DFAM), which focuses on enhancing discriminative power of a CNN and help alleviate issues caused by characteristics of remote sensing imagery. The third method introduced is Adaptive Task Aware Loss (ATAL), which focuses on improving the adaptive abilities of the network by leveraging information present in few-shot learning tasks.

This research has hypothesised that the methods introduced enhance the overall performance of the proposed network, providing a robust solution to challenges of CDFS scene classification where remote sensing imagery is utilised.

1.3 Research Aim and Objectives

1.3.1 Research Aim

The aim of this dissertation will be to focus on modifying an existing CNN to solve CDFS scene classification for a limited label data domain such as the remote sensing image domain. This will be achieved by improving the network's adaptability and robustness, which will enable it to handle novel remote sensing image classes in contexts where significant domain shift exists, and a limited amount of labelled data is available.

1.3.2 Research Objectives

A breakdown of the proposed framework, that will be used to achieve the aim is detailed below:

1. Develop and implement a cross-domain augmentation technique that reduces the domain shift, enhances the adaptation ability of a CNN and improves generalisation performance when classifying unseen remote sensing image classes.
2. Design an attention mechanism that increases discriminative power to further enhance the classification capability of a CNN, given the complex characteristics associated with remote sensing imagery.
3. Design a method to improve the adaptive ability of a CNN by effectively leveraging information from few-shot learning tasks during training to improve the loss function.
4. Develop a CNN that can effectively work with a limited amount of labelled remote sensing data and evaluate its overall performance in terms of prediction accuracy and ability.
5. Evaluate the efficacy of the proposed framework in comparison to traditional, related attention and augmentation methods in a CDFS setting with remote sensing data.
6. Achieve comparable CDFS scene classification performance when compared to existing and related approaches.

The three additions mentioned in section 1.2 are directly related to the objectives listed above and are essential to the main aim of this dissertation. They will be integrated and applied to achieve the objectives accordingly.

1.4 Research Methodology

A prototypical network serves as the base for the proposed method together with three modifications that were required and verified to enable the achievement of the main aim, stated in section 1.3.1. The dissertation introduces the modifications in one complete network that will be used in a CDFS setting, but each modification will be discussed separately as detailed below:

1. A data augmentation technique introduced as Cross-Domain Region MixUp (CDRM) will be discussed in section 3.6.2.
2. An attention solution introduced as a Dynamically Focused Attention Module (DFAM) will be discussed in section 3.6.3.
3. An adaptive loss introduced as Adaptive Task Aware Loss (ATAL) will be discussed in section 3.6.4.

Since these three modifications are combined into a single framework, it is important to verify the individual contributions of each method to the dissertation. Therefore, an ablation study is utilised to validate the efficacy of the network that is proposed and justify each addition.

1.5 Main Findings

The principal conclusions of this dissertation are as follows:

1. The cumulative effect of the three modifications, CDRM, DFAM and ATAL, that are introduced in this dissertation and combined into a single framework have demonstrated effectiveness in the CDFS classification tasks of remote sensing images. Furthermore, each of the additions has been validated and shows that they ultimately enhance the classification capability of the network.
2. CDAN outperforms traditional and related methods in CDFS classification tasks on remote sensing data. It was also observed that remote sensing dataset characteristics tend to influence the performance of the proposed and comparison networks in CDFS scenarios.
3. Supervised learning was found to be less effective in a low data availability setting. Another notable finding is that regular attention modules and augmentation methods were found to be less effective than expected in a CDFS setting, which was most likely due to possible overfitting on the source domain data and learnt features that were not domain invariant.

4. The main finding of the dissertation is that the complete proposed network, CDAN, is effective in a setting where a large domain shift exists between source (natural image dataset) and target datasets (remote sensing imagery). Additionally, it performed well in a low data availability scenario where only a few labelled samples were accessible.

1.6 Research Limitations and Assumptions

It is reasonable to introduce limitations and assumptions to ensure a manageable scope, given the complex nature of the research topic. The first limitation is that the benefits of CDRM, DFAM and ATAL are limited to CNNs and scene classification tasks of remote sensing data. Therefore, non-CNN and non-image classification domains are not considered. Secondly, CDFS learning for scene classification may not generalise well or adapt to domains that are not used in this dissertation. Lastly, dataset characteristics can influence the results that are achieved depending on how well the proposed network adapts to different datasets within a CDFS setting.

In terms of assumptions, these had to be in place before the experiments to ensure a reasonable and feasible scope. The first assumption is that the CDFS learning paradigm may be a suitable starting point for scene classification within a limited label data domain such as remote sensing. The second assumption is that the CDFS learning paradigm is suitable for handling domain shift, that occurs due to the differences between feature distributions of datasets. The third assumption is that CNNs serve as a reasonable starting point to introduce improved adaptation and generalisation within the CDFS learning paradigm. A detailed breakdown of limitations and the relevant assumptions will be provided in chapter 3 (section 3.8).

1.7 Summary

The remainder of the dissertation is structured with the following chapters as follows:

- Chapter 2 consists of a comprehensive literature review in the field of image classification. A diverse range of optical and remote sensing image classification approaches are analysed to enable a deeper understanding of this research and its origin.
- Chapter 3 provides information on the datasets that will be used and discusses the methodology to solve the research problem. It includes in-depth explanations of each contribution and presents the structure of the complete proposed network.
- Chapter 4 presents and discusses the results of the experiments as well as the ablation study. The experiments confirm the efficacy of the proposed network with the modifications, and the ablation study proves the value of each contribution to

the overall model. The experiments also serve to validate the hypothesis, which states:

> A CNN enhanced with advanced data augmentation and attention techniques can effectively address significant domain shifts and achieve comparable or superior performance to existing well-known methods in cross-domain few-shot scene classification tasks with a limited amount of labelled remote sensing data.

- Chapter 5 contains the conclusion. It outlines the contributions and a summary of the research. Finally, the limiting factors of the dissertation as well as possible improvements for further research are discussed.

Chapter 2

Literature Review

2.1 Introduction

Existing methods of scene classification and utilisation of remote sensing images in related research will be reviewed in this chapter.

Supervised learning methods are capable of achieving impressive results when performing scene classification tasks with remote sensing data [Chaib *et al.* 2017; Gong *et al.* 2018; Li *et al.* 2017a; Ma *et al.* 2019a; Guo *et al.* 2019]. However, supervised deep learning requires substantial quantities of labelled data to effectively train a network and recognise the classes within a dataset [Alajaji *et al.* 2020]. The size of remote sensing datasets is problematic as they typically contain a few hundred labelled images per class. Therefore, techniques such as transfer learning are often required to enhance model performance when using supervised learning approaches [Yu *et al.* 2017].

There are various existing methods that seek to boost the performance of CNNs, such as data augmentation, transfer learning or attention methods. Data augmentation is the process of modifying existing data to create additional data for training of models, which leads to enhanced generalisation ability of the model [Filipi Gonçalves dos Santos and Papa 2022; Zhang *et al.* 2017; Yun *et al.* 2019; DeVries and Taylor 2017]. Transfer learning with CNNs leverages the weights from a network trained on one task being used for another related task [Torrey and Shavlik 2010]. It is beneficial as it speeds up the learning process while reducing the amount of labelled data needed for training [Wang *et al.* 2018c]. Additionally, attention methods direct the focus of a CNN to the important regions within input data, therefore improving classification performance [Zagoruyko and Komodakis 2016].

Beyond existing supervised learning methods, advanced deep learning techniques have been created to perform scene classification on remote sensing data. This includes semi-supervised, unsupervised, and generative approaches. Semi-supervised methods

leverage unlabelled data to supplement the training process [Wang *et al.* 2019; Ren *et al.* 2018], while unsupervised approaches can be trained without any labelled data [Lu *et al.* 2017; Zhang *et al.* 2020b]. Alternatively, generative methods are able to create new labelled data to enhance classification performance [Ma *et al.* 2019b; Wang *et al.* 2019].

In response to data requirements of supervised learning, few-shot learning and CDFS learning have emerged as promising alternative approaches for performing scene classification of remote sensing images. Regular few-shot learning can be considered as a deep learning method that attempts to simulate the human ability to recognise novel categories/classes given a limited amount of samples. It is also known as N-way K-shot learning since N classes and K labelled samples are used per class during the training phase [Chen *et al.* 2019; Alajaji *et al.* 2020; Finn *et al.* 2017; Guo *et al.* 2020; Li *et al.* 2020ab 2021; Zhang *et al.* 2020a 2021; Vinyals *et al.* 2016]. Meanwhile, CDFS learning research is more recent and attempts to overcome challenges associated with regular few-shot learning by adapting models across different domains without the need for extensive re-training [Guo *et al.* 2020; Tseng *et al.* 2020].

2.2 Remote Sensing Imagery

The use of data from the remote sensing field has played a crucial role in environmental monitoring, change detection, urban flood modelling, urban water and gas pollution, urban planning and geographical mapping in relation to classification [Huang *et al.* 2019; Wang *et al.* 2018b; Tane *et al.* 2018]. This type of data has been used for decades, it began when satellites such as the Landsat 1 were fitted with sensors to collect large-scale, multi-spectral imagery and provided a new perspective of the Earth [Hemati *et al.* 2021].

Scene classification of remote sensing imagery refers to the process of assigning labels or land cover classes to entire scenes or regions within an image [Al-Doski *et al.* 2013]. Advancements in remote sensing technology have significantly increased the volume of available information in this domain over time, which enables improved usage and detailed analyses of the data. Furthermore, simple digital processing techniques or manual interpretation were common in the past, but the use of modern techniques such as machine learning have enhanced efficiency and accuracy for these tasks [Hemati *et al.* 2021; Al-Doski *et al.* 2013].

Despite the improvements brought by the introduction of machine learning, classification tasks remain challenging given the properties of remote sensing images. This includes a high resolution, variance of the images in scale and orientation [Mekhalfi *et al.* 2015] in addition to intricate spatial relationships between entities within a scene [Li *et al.* 2020b]. These characteristics cause high intra-class variance, high inter-class similarity and increased computational costs for data processing, which hinders the effectiveness and efficiency of model performance [Ma *et al.* 2015; Guo *et al.* 2019; Li *et al.* 2020b] as discussed in section 1.1.

Instances of elevated intra-class variance and significant inter-class similarity can be observed in figure 2.1.

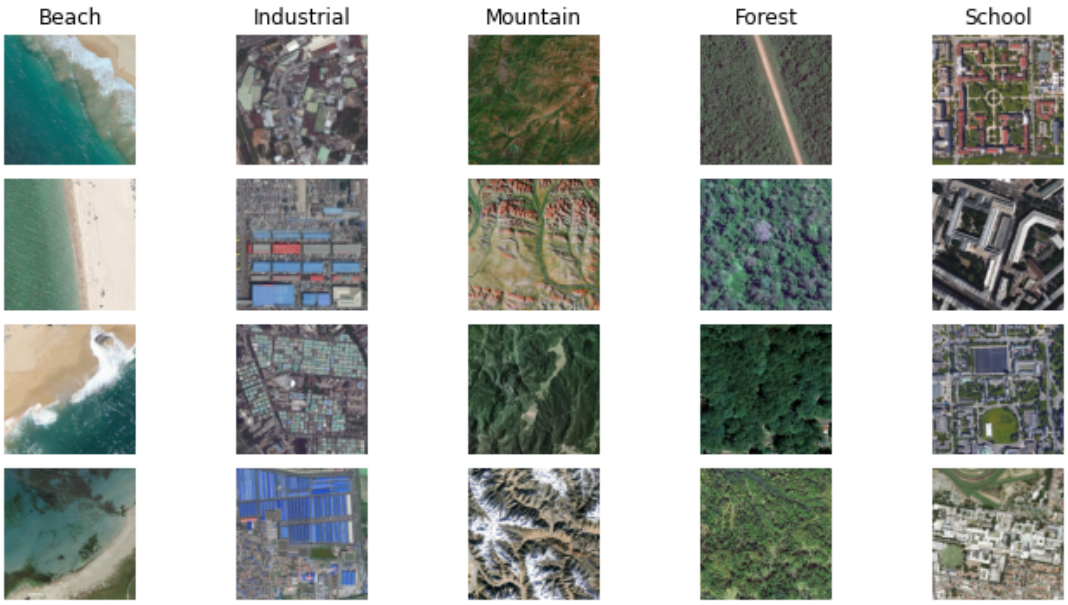


Figure 2.1: Samples from multiple classes in the Aerial Image Dataset.

2.3 Convolutional Neural Networks

2.3.1 Nature of CNNs

CNNs are regarded as a fundamental tool within the computer vision domain as they have significantly improved the analysis of image datasets [Hossain and Sajib 2019]. One of the earliest breakthroughs with CNNs was achieved with AlexNet in the ImageNet competition, which demonstrated the capability of deep learning in performing different functions with large-scale datasets [Krizhevsky *et al.* 2012] such as classification/recognition. CNNs are designed to automatically learn hierarchical features at different levels, this makes them effective for image-related tasks.

The fundamental term that defines this type of network is *convolution*. In CNNs, convolutional filters are used to traverse an input image and then compute the weighted aggregate of values for pixels within its receptive field. The mathematical definition of *convolution* at a specific pixel location (i, j) is described as:

$$C(i,j) = \sum_{m=-a}^{a} \sum_{n=-a}^{a} I(i+m, j+n) \cdot K(m,n), \tag{2.1}$$

where $C(i, j)$ represents the resultant matrix of a convolution operation at pixel position (i, j). K represents the convolutional filter being applied, $K(m, n)$ represents the filter value at (m, n), I indicates the original image and $I(i+m, j+n)$ signifies the value of a pixel within the image at the $(i + m, j + n)$ coordinates. The filter moves across an image in a sliding-window fashion with a fixed step-size is known as the stride. The resultant matrix produced from the convolution operation described above is a feature map, which captures a response when the filter is applied to different parts of an image and detects important features within the image.

In CNNs, filters at different layers capture different levels of abstraction in the input image. The filters that are located closest to the input layer will typically capture and learn basic attributes such as lines and edges, whereas filters that are found within deeper layers can capture intricate features, which includes shapes, textures and colours [Zeiler and Fergus 2014]. Mathematically speaking, the equation (2.1) is technically called *cross-correlation*, therefore the term is used interchangeably with *convolution* as the rotation of the filter by 180 degrees has no effect on the CNN's ability to learn [Bengio *et al.* 2017]. It is also essential to acknowledge that both operations (convolution and cross-correlation) provide *translational equivariance*, which ensures the detection of features by the network, regardless of the spatial positions of features within the input space.

Another important aspect to understand in the application of CNNs is that the filter operates across all channels simultaneously when the convolution operation is applied on multi-channel images. The spatial dimensions of a filter are only referred to when describing the convolution operation; this is because it only produces one feature map across all of the channels which is then summed to produce a feature map. For example, consider an RGB image, each filter would span the three colour channels (red, green, and blue), and the output feature map would be computed by summing the contributions from each channel. Distinct filters are used for each channel which means that they each have their own set of weights. These weights constitute the set of trainable parameters that are modified to approach a local minimum based on their influence on the overall loss of the network. This optimisation is achieved through the process of *backpropagation* and *stochastic gradient descent*.

2.3.2 Attention Mechanisms

Attention in CNNs helps improve classification ability by guiding networks to focus their attention on the correct regions of the input imagery [Zagoruyko and Komodakis 2016]. The introduction of different variants of attention mechanisms within CNNs to perform classification tasks has garnered significant interest, as they are able to leverage the channel and spatial information within training data to improve model performance in different types of datasets. Spatial information indicates "where" the network should focus on whereas spectral/channel information indicates "what" the feature is in an input image [Woo *et al.* 2018].

The usage of attention modules/mechanisms first became popular in research involving natural image datasets. Early research such Wang *et al.* [2017] proposes the use of a residual attention model that consists of attention modules that are stacked together which is able to create attention-aware features and ensures good performance for classification tasks. Furthermore, the introduction of a convolutional block attention module (CBAM) that is lightweight and can be easily integrated into CNNs led to more robust features being learnt Woo *et al.* [2018]. The modules seek to derive information from the spatial and channel dimensions to create feature maps, which then undergo multiplication with input data for adaptive refinement of the features. The attention module consists of two components connected in series. Firstly, the global attention module focuses on channel information and utilises information from all channels. Secondly, a local attention module focuses on the spatial information at a single scale. Figure 2.2 depicts how the CBAM is structured and how it integrates into a CNN.

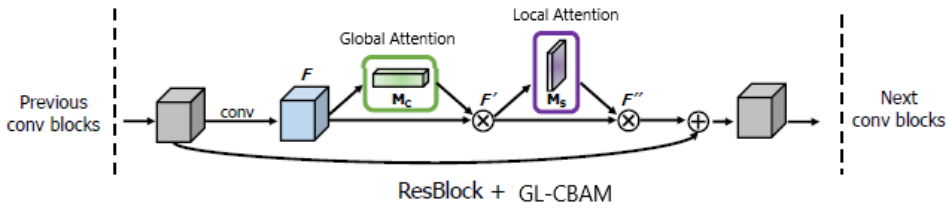


Figure 2.2: A visualisation of how CBAM is set up between ResBlocks [Woo *et al.* 2018].

Application of attention modules/mechanisms in research with mainstream remote sensing data are numerous. They have been proven to be effective in increasing the discriminative power of networks and alleviating issues associated with remote sensing imagery such as intra-class variance and inter-class similarity [Guo *et al.* 2019; Yuan *et al.* 2020; Li *et al.* 2020b]. Guo *et al.* [2019], Yuan *et al.* [2020] and Li *et al.* [2020b] all utilise global and local attention mechanisms for the classification of aerial imagery, their networks contain a feature extractor with mechanisms that capture global and local semantic features by using attention mechanisms in a similar manner to Woo *et al.* [2018]. Furthermore, the work of Kong *et al.* [2021] improves upon the original CBAM with the addition of a squeeze-excitation module to the existing structure, but arranges the modules in parallel and proves it is effective for scene classification on remote sensing imagery. The use of attention modules discussed above and their variants such as cross-attention mechanisms [Li *et al.* 2020d], stand-alone channel attention mechanisms [Tong *et al.* 2020] and class-wise attention mechanisms [Wang *et al.* 2023b] shows that attention methods remain a popular choice to boost the performance of models for classification tasks on remote sensing imagery.

Interesting research known as channel gating/pruning has also been carried out on neural networks to improve performance. The work of Gao *et al.* [2018], Hua *et al.* [2019] and Ehteshami Bejnordi and Krestel [2020] explores channel gating/pruning, which is the practice of reducing the computational load on CNNs by only focusing on relevant channels during the training process, which also helps networks focus on important features while maintaining performance.

The research discussed in this section shows that attention mechanisms/modules have a significant impact on classification performance for natural image and remote sensing data. Therefore, in this dissertation, the use of attention modules will be explored while utilising ideas related to channel gating/pruning for scene classification tasks. The upgraded attention modules aim to advance the current work on attention mechanisms/modules and will be discussed in section 3.6.3.

2.3.3 Metric Learning

Approaches that utilise this type of learning method are effective in scene classification tasks. It involves learning a distance function, which is mainly accomplished by assessing the resemblance or dissimilarity between two objects [Weinberger and Saul 2009]. In the case of image classification, metric learning helps a model with category/class assignment for a test image by identifying a category/class that it shares a minimum distance with based on the metric used. This technique is advantageous in few-shot contexts, where there is restricted number of labelled samples or data points available. There are popular few-shot methods with metric learning that include Chen *et al.* [2019]; Vinyals *et al.* [2016]; Snell *et al.* [2017]; Sung *et al.* [2018] and they use Euclidean distance or the Cosine similarity metric for scene classification purposes. Interestingly, Cosine similarity is a metric that has been proven to assist with challenges such as high intra-class variance and high inter-class similarity [Wang *et al.* 2018a; Xu *et al.* 2021b].

The use of metric learning in models that focus on remote sensing imagery for scene classification tasks are abundant. Firstly, Gong *et al.* [2018] employs a method known as deep structural metric learning with a technique that promotes diversity. The deep structural metric learning extracts additional structural information regarding features of images during the training phase to increase classification ability, while the diversity-promoting method has a regularising effect on the network. Additionally, Cheng *et al.* [2018] proposes a network to help alleviate remote sensing image classification challenges, such as intra-class variance and inter-class similarity, through the use of a discriminative CNN. The network includes an improved discriminative objective function with a metric learning related regularisation expression for increased performance.

Furthermore, a range of research that uses remote sensing imagery with a few-shot learning based approach are also popular for scene classification tasks. Li *et al.* [2020b] propose a matching network for few-shot learning called DLA-MatchNet, which com-

bines discriminative learning and metric learning into an adaptive network. Attention mechanisms perform discriminative learning while an adaptive matcher acts as a flexible learning metric. Another interesting network proposed by Zhang *et al.* [2020a] introduces an effective few-shot classification method that utilises the EMD metric. The EMD metric was initially utilised for image retrieval, which is used in the calculation of distance between structural features [Rubner *et al.* 2000]. In other words, the EMD provides optimal matching flows that exist between image features of minimal distance from each other. The feature extractor in Zhang *et al.* [2020a] passes support and query set embeddings to an EMD cross-mechanism, which is responsible for weight assignment according to feature importance. Research by Yuan *et al.* [2020] improves upon the EMD network with the addition of global and local attention mechanisms in their feature extractor, which results in increased discriminative power of the model on the remote sensing images and leads to improved few-shot classification. The only downside of EMD is that it is computationally expensive.

There are a wide variety of approaches that use variations of metric learning for scene classification, as discussed in this section. This research will employ the use of a simple, efficient and effective metric such as the Cosine similarity metric. It should be noted that metric learning is not a major area of focus in this research but is the chosen metric for classification.

2.3.4 Loss Functions

A loss function is an essential component of a network that is utilised during the training process. The main aim of a loss function is to assess accuracy of a network's predictions in relation to actual labels of the samples. The function is able to measure this discrepancy and then guides the optimisation process to minimise the discrepancy, which improves prediction accuracy [Wang *et al.* 2020]. There are a number of loss functions and each offers different advantages.

Firstly, cross-entropy loss is widely utilised in networks that focus on classification, which is due to its effectiveness [Chen *et al.* 2019; Zhang *et al.* 2020a; Cheng *et al.* 2018; Li *et al.* 2020c; Ma *et al.* 2019a; Snell *et al.* 2017; Vinyals *et al.* 2016]. In simple terms, this loss function can measure network performance where output is in the form of a prediction probability that has a value between 0 and 1. It is able to help differentiate the predicted class probabilities from the actual distribution to produce more reliable results and is therefore highly effective at inter-class separation [Wang *et al.* 2020]. The advantages of this loss function are numerous, it is effective in multi-class scenarios [Demirkaya *et al.* 2020], has smooth optimisation [Berrada *et al.* 2018] and is flexible in its utilisation [Kim *et al.* 2021]. One of the only drawbacks of cross-entropy loss is that when the training and testing data distributions are significantly dissimilar its effectiveness is reduced [Xu *et al.* 2021a], which is not ideal in a cross-domain problem setting.

Secondly, the focal loss function is used for classification tasks and is designed to handle class imbalance scenarios within datasets. This is achieved with a modified cross-entropy function by placing emphasis on difficult cases [Lin *et al.* 2017]. Lastly, a loss function known as Mean Squared Error (MSE) is used for regression tasks but can be effectively adapted for classification purposes. This is achieved by calculating and comparing the average squared differences between the actual and predicted values. [Li *et al.* 2020bd; Sung *et al.* 2018].

A focus area in this dissertation will be loss function enhancement. The Cross-entropy loss will serve as base and starting point, due to its popularity and effectiveness as discussed in this section. The focus will be centred on improving the function to operate effectively in a context that focuses on CDFS learning. This will be discussed in section 3.6.4.

2.4 Learning Paradigms

2.4.1 Preliminary

A brief explanation of classification for the relevant learning paradigms (supervised, few-shot and CDFS classification) will be discussed with notation before the overall literature in this section is introduced.

2.4.1.1 Standard Supervised Classification

Standard supervised classification assumes that substantial amounts of labelled data will be readily available for training. The model learns from the data by extracting patterns and relationships to be able to generalise effectively on unseen data. It is essential to acknowledge that training and testing datasets contain samples from the same classes in this approach [Liu *et al.* 2018; Ma *et al.* 2019a; Li *et al.* 2020c].

The dataset is defined by $\mathbf{D} = \{\mathbf{D_{Train}}, \mathbf{D_{Val}}, \mathbf{D_{Test}}\}$. The training set consists of labelled pairs that are defined as $\mathbf{D_{Train}} = \{(\mathbf{x_i}, \mathbf{y_i})\}_{\mathbf{i=1}}^{\mathbf{N}}$, $\mathbf{y_i} \in \{\mathbf{1}, ..., \mathbf{C_{total}}\}$, where $\mathbf{N}$ stands for the number of samples and $\mathbf{C_{total}}$ represents the number of classes within $\mathbf{D_{Train}}$. The validation set $\mathbf{D_{Val}}$ also consists of labelled pairs which are identical to the ones found within the training set. It is utilised to assess model performance throughout the training phase. The validation data is extremely important for hyperparameter tuning and model adjustments to reduce the chance of overfitting.

The test set $\mathbf{D_{Test}}$ can be defined as $\mathbf{D_{Test}} = \{(\mathbf{x_m})\}_{\mathbf{m=1}}^{\mathbf{M}}$, where $\mathbf{M}$ stands for the number of samples. The test set consists of labelled samples, and labels are only used to assess the model's performance after the training and validation phases are complete. The test set is crucial in obtaining an unbiased evaluation of the model that has been trained and validated. In summary, a a model $\mathbf{f}_\theta(\mathbf{x})$ is parameterised by network pa-

rameters θ, is trained on $\mathbf{D_{Train}}$, validated on $\mathbf{D_{Val}}$ and then tested on $\mathbf{D_{Test}}$ to predict labels $\widehat{\mathbf{y}}$ for each sample $\mathbf{x_m}$ in $\mathbf{D_{Test}}$. The predicted label $\widehat{\mathbf{y}}$ for a sample will be a class in $\{\mathbf{1}, ..., \mathbf{C_{total}}\}$.

2.4.1.2 Few-Shot Classification

This approach for classification is designed to perform deep learning and make predictions based on a limited amount of data. The training procedure is commonly called meta-training, while the testing procedure is known as meta-testing within this approach/field. The dataset is referred to as a meta-dataset and may be defined as $\mathbf{D} = \{\mathbf{D_{base}}, \mathbf{D_{val}}, \mathbf{D_{novel}}\}$. It has the same structure as the supervised sets above, but the difference is where $\mathbf{C}$ defines the classes in this dataset and $\mathbf{C_{base}}, \mathbf{C_{val}}, \mathbf{C_{novel}}$ are mutually disjoint [Vinyals *et al.* 2016].

Few-shot classification involves task-based learning which enables a network to correctly classify novel classes that were not available during the training process. It is also accomplished with a limited quantity of labelled samples. Therefore, the approach tests a network's adaptability when presented with new information. Each task has an N-way K-shot structure, where $\mathbf{N}$ represents the number of classes and $\mathbf{K}$ represents the number of samples. A task consists of:

- Support Set ($\mathbf{S_i}$): It follows an N-way K-shot structure, which means it has $\mathbf{K}$ samples from each of the $\mathbf{N}$ classes. Therefore, it consists of $\mathbf{N} \times \mathbf{K}$ labelled samples.
- Query Set ($\mathbf{Q_i}$): The aim of this set is to complement the support set within each task to test the network's capacity to correctly classify unlabelled samples based on what it has learnt from the support set.

The tasks/episodes for meta-training and meta-testing have the same structure in the few-shot approach [Zhang *et al.* 2021]. A visual example of few-shot training and testing can be seen in Figure 2.3. Ultimately, the aim is to learn a model g from $\mathbf{D_{base}}$, with $\mathbf{D_{val}}$ being utilised for validation during meta-training, and to enable the model to accurately classify novel/unseen classes from $\mathbf{D_{novel}}$ with a limited amount of samples to learn from during meta-testing. The number of samples can range from 1 to 10. Therefore, this process allows the network to adjust quickly when encountering new tasks with the least quantity of data which enhances efficiency and adaptability.

2.4.1.3 Cross-Domain Few-Shot Classification

CDFS classification extends the idea of regular few-shot classification for data-limited domains. It also addresses issues such as differences in training and testing data distributions, which will be discussed in section 2.4.5.1. The main CDFS category that will

be used in this dissertation is the single source category. It uses a single dataset from a domain with more data availability in meta-training, which is known as the source dataset ($\mathbf{D_{Source}}$). This approach enables a network to learn features that are more generalised in nature. A dataset that originates from a data-limited domain is then used in meta-testing and is known as the target dataset ($\mathbf{D_{Target}}$). The tasks/episodes for meta-training and meta-testing processes in CDFS classification have a similar structure to the tasks associated with regular few-shot classification.

Meta-Training and Meta-Testing

CDFS learning with meta-training and meta-testing is instrumental in solving the research problem as it enables the network to perform classification with a limited amount of labelled data, which will be discussed in section 2.4.5. It is essential for limited labelled data domains like remote sensing imagery that will be used in this dissertation. Furthermore, it is important to note that the few-shot learning applied in a cross-domain setting with the proposed network is classified under the single source category where the target domain data is accessible, as described in section 2.4.5. This is because sampling of data from the target domain occurs and is mixed with the source domain data during training.

The procedure for CDFS network learning can be formalised as follows: the meta-training phase utilises the base source dataset $\mathbf{D}^{\mathbf{base}}_{\mathbf{Source}}$, which is a large dataset with numerous classes, in addition to randomly sampling a limited amount of data from the base target dataset $\mathbf{D}^{\mathbf{base}}_{\mathbf{Target}}$. During meta-training, for each task/episode, an N-way K-shot structure is used. The support set is defined as $\mathbf{S} = \{(\mathbf{X_s}, \mathbf{Y_s})\}$, whereas the query set is defined as $\mathbf{Q} = \{(\mathbf{X_q}, \mathbf{Y_q})\}$. Both sets are generated through sampling $\mathbf{D}^{\mathbf{base}}_{\mathbf{Source}}$ in addition to random samples from $\mathbf{D}^{\mathbf{base}}_{\mathbf{Target}}$. The meta-testing phase will mirror the meta-training, but a support and query set will be made up of the novel classes found in $\mathbf{D}^{\mathbf{novel}}_{\mathbf{Target}}$. Therefore the model observes a support set $\mathbf{S}^{\mathbf{Target}} = \{(\mathbf{X}^{\mathbf{Target}}_{\mathbf{s}}, \mathbf{Y}^{\mathbf{Target}}_{\mathbf{s}})\}$ and will predict labels for the query set $\mathbf{Q}^{\mathbf{Target}} = \{\mathbf{X}^{\mathbf{Target}}_{\mathbf{q}}\}$.

2.4.2 Supervised Learning

The use of supervised approaches with deep learning necessitates substantial quantities of labelled data to solve classification problems, but still remain a popular choice as seen in research papers over the past eight years [Gong *et al.* 2018; Chaib *et al.* 2017; Liu *et al.* 2018; Ma *et al.* 2019a; Cheng *et al.* 2018; Guo *et al.* 2019; Li *et al.* 2020c; Zeng *et al.* 2021; Kong *et al.* 2021].

There are interesting methods that were established within the realm of supervised learning to reduce reliance on extensive datasets. Firstly, there is transfer learning which entails the use of weights from a network trained on a specific task being used for another related task [Torrey and Shavlik 2010]. It is beneficial as it helps net-

works learn faster and reduces the data requirements for the training process [Wang *et al.* 2018c]. The work of Li *et al.* [2020c] investigates and implements the use of transfer learning, with weights from networks trained on ImageNet2015, which is then applied to deep CNNs for further training and fine-tuning to classify high-resolution remote sensing images and achieve better results. Additionally, further research with supervised approaches such as Shang *et al.* [2018] utilises a deep memory CNN, which is implemented with the task of remembering spatial features. A transfer parameter technique is also employed to aid with convergence and reduce the required training time.

2.4.3 Unsupervised Learning

Unsupervised learning has gained popularity as an alternative to supervised learning, especially in domains where large labelled datasets are hard to acquire [Lu *et al.* 2017; Zhang *et al.* 2020b]. Unsupervised methods are complex but aim to learn spatial and spectral structures as well as patterns within remote sensing imagery. The meaningful feature representations produced are useful when performing classification [Lu *et al.* 2017].

Firstly, Lu *et al.* [2017] presents a complex methodology that executes remote sensing classification with an unsupervised representation method that involves three components. The first is a deconvolution network for feature learning, which minimises a reconstruction error between input data and its deconvolved result. A spatial pyramid is then used for feature aggregation at multiple scales, followed by the generation of discriminative representations for input data through the combination of output from the spatial pyramid and feature extractor. The resultant representation vector serves as input for a support vector machine which finally performs the classification.

There are also other interesting approaches that make use of unsupervised learning and combine it with other approaches to improve their performance. Chen *et al.* [2018] proposes a network which consists of three parts that combine unsupervised learning with supervised learning. The first part utilises unlabelled data and extracts salient regions while the second part learns features that are present in the salient regions using sparse filters. The learnt parameters are then used for initialisation of the CNN and the third part of the network undergoes fine-tuning on the labelled data. Additionally, Zhang *et al.* [2020b] seeks to enhance the classification of polarimetric synthetic aperture radar (PolSAR) imagery by using an unsupervised deep learning method combined with few-shot learning. A contrastive learning network designed to handle the data undergoes unsupervised pre-training, and the feature representations are used in few-shot learning classification tasks.

2.4.4 Generative Learning

Unsupervised learning is useful for extracting features from unlabelled data but struggles to achieve the performance seen with the use of supervised methods [Yan *et al.* 2020]. Generative learning offers an alternative approach by generating additional realistic samples for training models. It is a powerful method that has gained a lot of attention to address issues related to limited labelled datasets [Singh and Bruzzone 2021; Ma *et al.* 2019b 2022]. A few of these interesting approaches that perform classification on remote sensing imagery are explored below.

Firstly, the work of Ma *et al.* [2019b] proposes a generative adversarial method that generates substantial quantities of labelled samples for the data augmentation process. Their methodology has three main components. The first component is a traditional generative adversarial network with an Online-Output method. This technique generates samples during the training process and constantly uses models to generate samples to ensure diversity. The second component is Generative-Model-Sifting which then selects the best model according to loss values. Sifting in the context of this paper means that the models and their output are evaluated, only the diverse and high-quality output models are retained. The third component, known as Labelled-Sample-Discriminating, acts as quality control and removes unwanted samples to work with the rest of the network in performing effective classification.

Secondly, Wang *et al.* [2019] proposes a novel deep generative model, variational generative adversarial network (VGAN), that can work with a limited labelled hyperspectral image dataset and additional unlabelled data. Therefore, the VGAN is adapted to operate within a semi-supervised setting. An ensemble prediction technique is used to leverage the unlabelled data to enhance network performance, while a classifier network learns the relationship between the latent space and true image space. A cooperative relationship is formed between the classifier and VGAN to help generate relevant samples that enhances classification ability.

Lastly, Singh and Bruzzone [2021] proposes a generative adversarial network (GAN) for synthetic multi-spectral image generation. The proposed network leverages spectral indices, that capture class-specific spectral properties and guide the network in generating realistic class-discriminative images. Therefore, the variety of training data will be improved, which enhances network classification performance on multi-spectral remote sensing tasks. The method demonstrates better performance compared to traditional GAN approaches.

As discussed in this section, generative learning approaches from remote sensing imagery are powerful and effective for creating realistic labelled samples in data-limited scenarios. However, it should be noted that they do have complex setups with sophisticated models like GANs, which may lead to increased implementation difficulty and computational costs.

2.4.5 Few-Shot Learning

Few-shot learning has emerged as an approach that can classify examples from an unseen category/class with a limited amount of labelled data. It also assumes that data in the training and testing set have a similar distribution [Wang *et al.* 2023a]. It is also a viable alternative to generative approaches, which focus on creating labelled samples for training, while few-shot approaches aim to learn good representations from the limited amount of available data and are less complex.

Few-shot learning approaches usually adopt a meta-learning framework which aids in distinguishing it from other methods. Meta-learning can also be described as "learning to learn". This technique allows the training procedure to mirror the testing procedure across learning tasks. A key aspect of this approach is that a network is trained across a series of tasks, instead of one large dataset, which improves a network's ability to adapt when it encounters new, unseen tasks.

This learning approach enables the network to effectively generalise on new or unseen tasks that it encounters. A network is conditioned during meta-training to adapt to tasks and generalise well. During meta-testing, the structure of the tasks mirrors the meta-training procedure, but novel class data is used to evaluate the network's classification ability. The overarching approach remains the same in few-shot learning, but approaches may differ across frameworks. A representation of few-shot classification is visualised in Figure 2.3 and compares a supervised learning setup for classification to the few-shot setup. A variety of prominent few-shot networks have influenced existing research, and each of the networks will be discussed in this section.

Firstly, the work of Vinyals *et al.* [2016] proposes a matching network that maps the support set sample and input sample onto an embedding space. Similarity computations are then performed by a differentiable k Nearest Neighbour (kNN) classifier, which utilises the Cosine similarity metric. The matching network has three important components. Firstly, a feature extractor learns embeddings from the support and query set, these are then mapped onto a high-dimensional embedding space. Secondly, an attention mechanism computes the importance of support set samples in relation to the resemblance with the query sample. Thirdly, a bidirectional Long Short-Term Memory (LSTM) is utilised for learning the probabilistic distribution of support set data. The LSTM sequentially processes embeddings to enhance the learnt representations by identifying dependencies within the support set. Therefore, during training this network is able to classify query samples by effectively utilising information learnt from the support data. The similarity metric quantifies the distance between the embeddings of support samples and the query sample, while the nearest-neighbour classifier computes the final prediction.

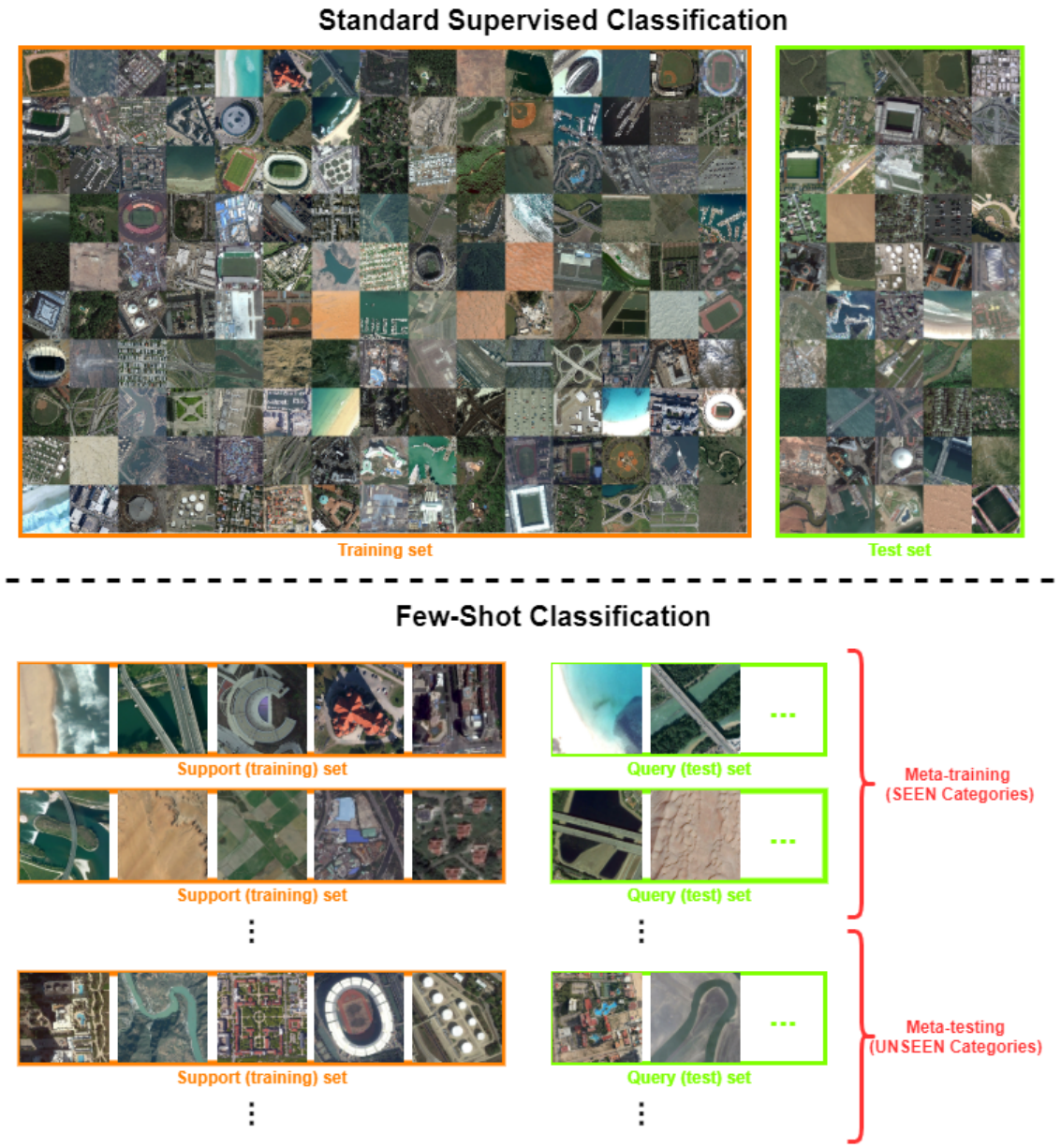


Figure 2.3: Comparison of dataset use in supervised classification vs few-shot classification. Adapted from Zhang *et al.* [2021].

Secondly, Snell *et al.* [2017] proposes a prototypical network (ProtoNet) for few-shot learning. A feature extractor (CNN) maps support and query set samples onto an embedding space. The network then computes prototypes for each category/class in the support set. Euclidean distance is then utilised to determine how close a query sample is to each class prototype within the embedding space. Lastly, a softmax classifier is utilised to determine the most probable class that the query sample belongs to. The work of Ren *et al.* [2018] improves upon ProtoNet to enable it to perform semi-supervised few-shot learning by accepting unlabelled examples during training for better cluster assignment.

Thirdly, Finn *et al.* [2017] proposes a model-agnostic method that focuses on improved initial parameter initialisation to decrease the required amount of gradient steps to achieve convergence. This is accomplished through the method utilising distributions of previous tasks to learn new tasks, which ensures good parameter initialisation and reduces the amount of data required for learning.

Lastly, Sung *et al.* [2018] propose a relation network. It comprises two major parts, the feature extractor and relation module. The first component is used to embed the support set samples and query set samples into a unified feature space. The second component, which is a neural network in nature, then takes the query image and support set output to compute a similarity score. Therefore, the similarity scores are calculated between the query image and each support image, with the highest score indicating the class to which the query image is assigned.

Additional research includes few-shot learning applied to remote sensing imagery that aims to improve on work introduced in this section. The work of Zhang *et al.* [2021] proposes a few-shot method for aerial imagery classification. The network has a feature extractor and utilises the Cosine distance metric that has an adjustable scaling parameter for classifier optimisation. The work of Li *et al.* [2020a] proposes a few-shot meta-learning method with a similar structure to Zhang *et al.* [2021], but Li *et al.* [2020a] also introduces an improved loss function called Balance loss that combines cross-entropy and a generalisation loss. This upgraded loss function aims to enhance generalisation ability by maximising the distance between different classes, which increases the linear separability. Other noticeable research that uses few-shot approaches for classification purposes with remote sensing imagery includes Zhang *et al.* [2020a], Li *et al.* [2020b] and Yuan *et al.* [2020]. These methods were also discussed in section 2.3.2 and section 2.3.3 as they utilise a combination of methods.

2.4.5.1 Cross-Domain Few-Shot Learning

Few-shot learning has gained popularity given its ability to recognise new classes with limited quantities of labelled data. In few-shot learning, it is presumed that the training and testing data will have similar feature distributions, which may be unrealistic in real-world scenarios. The difference in distribution in these scenarios is described as domain shift or domain gap. Regular few-shot learning does not account for this,

which has led to an increase in CDFS research that seeks to address this complex but realistic issue [Wang *et al.* 2023a].

CDFS learning is a subset of regular few-shot learning. It is classified into two main categories, multiple source and single source, depending on how the networks are trained. Multiple source means that multiple source domain datasets are used to capture different distributions to solve the CDFS problem. In this category one model or multiple models are used during the training process. Single source implies that only one source dataset is used to solve the CDFS problem, and in this category the network can either have access to sample some target domain data or is forbidden from accessing it during training [Wang *et al.* 2023a]. Figure 2.4 visualises the classification of the different CDFS approaches.

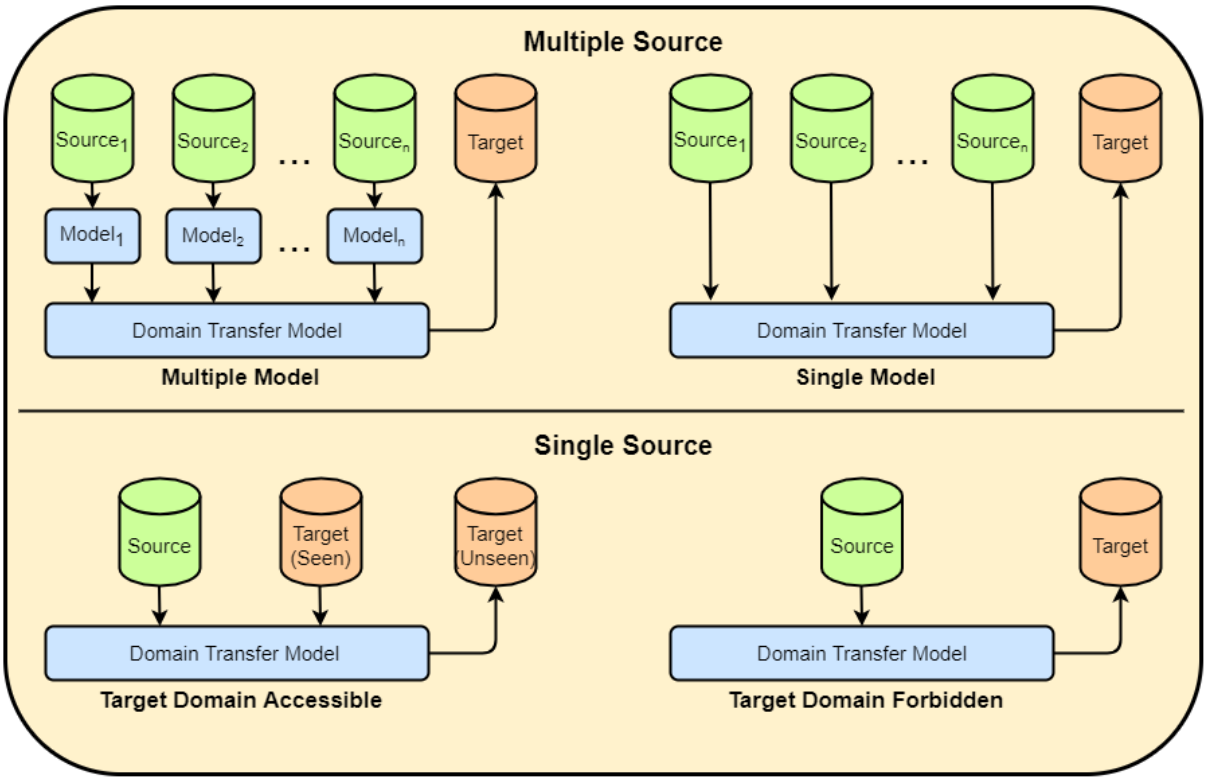


Figure 2.4: Comparison of existing types of CDFS learning approaches, which are split into multiple and single source categories. Adapted from Wang *et al.* [2023a].

CDFS learning is useful when working with data-limited domains such as remote sensing. A source dataset which is large and easy to obtain, as well as a target dataset that is limited in size, are mapped onto a common embedding space.[Guo *et al.* 2020]. Given the domain shift, the aim of CDFS learning is to have a network learn from the source dataset so that it can transfer knowledge and recognise new classes, with a few labelled samples from the target dataset [Li *et al.* 2021; Guo *et al.* 2020; Chen *et al.* 2019]. The single source CDFS method and embedding space is visualised in Figure 2.5. There are

a number of CDFS networks that use remote sensing imagery as input and concentrate on classification tasks. This type of research will be introduced and discussed below, but it is essential to acknowledge that the few-shot learning approaches presented in this section are equally applicable in cross-domain contexts, as demonstrated in Guo *et al.* [2020].

An initial research paper by Rostami *et al.* [2019] proposes a method that aims to facilitate knowledge transfer from an abundant Electro-Optical domain data to Synthetic Aperture Radar (SAR) domains. The embedding space that the datasets share is made domain-invariant, which enables knowledge to be transferred between domains and leads to enhanced generalisation ability with target domain data. Secondly, Tseng *et al.* [2020] proposes a CDFS based approach aimed at reducing domain shifts with layers that perform feature-wise transformation, which augments input image features during the training phase. This allows for the simulation of different feature distributions to enhance the robustness of learned representations towards domain shifts. This paper also considers adaptation from multiple domains to solve a CDFS problem, which involves training with multiple source datasets and generalising well during testing on one target dataset.

Additionally, Guo *et al.* [2020] attempts to study the effectiveness of CDFS learning across domains that are vastly different. They train with a fixed source domain and multiple target domains that have diverse characteristics, this includes a remote sensing dataset. In other words, they observe the effects that the domain shift has on classification effectiveness. Another interesting method to improve adaptability employs a deep cross-domain method to perform classification on remote sensing images. It utilises conditional adversarial domain adaptation and enables the network to overcome any domain shift due to the data, more importantly, alignment of the domain distributions is effectively achieved with this strategy Li *et al.* [2021].

Lastly, there are approaches that use meta-learning frameworks but employ unique techniques to solve CDFS problems. Ji *et al.* [2022] proposes integration of auxiliary tasks such as rotation prediction and contrastive learning to improve the transfer of knowledge across classes. This leads to better intra-class consistency as well as inter-class distinctiveness. Another interesting approach by Li *et al.* [2023] proposes three additions to enhance network performance for scene classification tasks that involves remote sensing imagery. Firstly, a Mean-teacher architecture is used to combine pseudo-labelling and consistency regularisation for better cross-domain learning, then dynamic thresholding to classify data into positive and negative samples for improved knowledge transfer. Lastly, a self-entropy loss is employed for uncertainty reduction and optimisation.

Given the research discussed in this section, CDFS approaches with a meta-learning framework are extremely powerful methods in research to address limited data availability challenges and domain shift in the remote sensing field. A majority of the methods are also simple to set up, efficient and available online, which reduces the implementation difficulty, but the research in this field is still maturing. A limited amount

of research exists for remote sensing scene classification that addresses domain adaptation therefore, solving this CDFS problem will be focused upon in this dissertation.

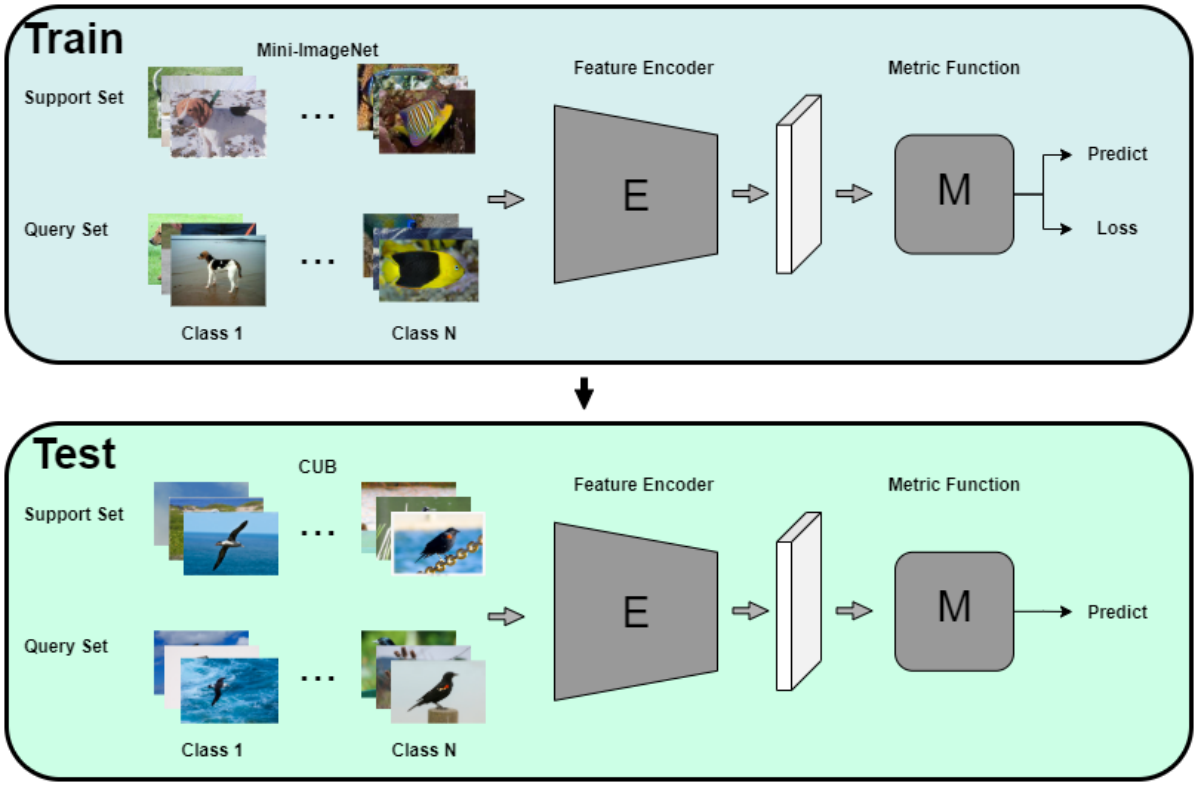


Figure 2.5: Single source CDFS learning visualisation of the training and testing procedure with a network. Adapted from Sa *et al.* [2022].

2.5 Data Augmentation

Data augmentation techniques are utilised to alter input data for a network during the training phase, with the aim of improving the data variation to increase the network's generalisation capability. [Filipi Gonçalves dos Santos and Papa 2022; Zhang *et al.* 2017; Yun *et al.* 2019; DeVries and Taylor 2017]. It is useful for datasets with a limited amount of labelled samples, as it allows for artificial expansion of the dataset without the need for additional data collection and labelling. Specifically, it is advantageous when dealing with remote sensing imagery or few-shot learning tasks, as it encourages a network to learn more generalised representations. Therefore, it leads to better feature extraction and improved network performance during testing [Sun *et al.* 2021].

Advanced data augmentation methods also exist to further improve quality of learnt representations and performance of vision networks. These include methods such as Cutout and Random Erasing, that remove or occlude arbitrarily chosen areas of an input image and force the network to focus on different regions [DeVries and Taylor 2017;

Zhong *et al.* 2020]. Other methods such as Mixup, CutMix and Cross-domain CutMix employ more creative techniques by blending or mixing input image data. These processes create new samples for training, improves data variability and makes the network more robust to ambiguous input [Zhang *et al.* 2017; Yun *et al.* 2019; Nakamura *et al.* 2022]. These advanced augmentation methods will be discussed more in this section.

2.5.1 Cutout

A simple but effective data augmentation that was developed is known as Cutout [DeVries and Taylor 2017]. The method involves the random removal/masking of square-shaped regions from the input image data before it is used during the training process. This approach compels a network to concentrate on surrounding areas within an image, which leads to more generalised representations being learnt. The act of removing random parts of the image can be seen as a form of regularisation and helps prevent overfitting. This method is easy to implement, introduces minimal computational overhead and may be integrated with various augmentation methods. The Cutout method is visualised in figure 2.6.

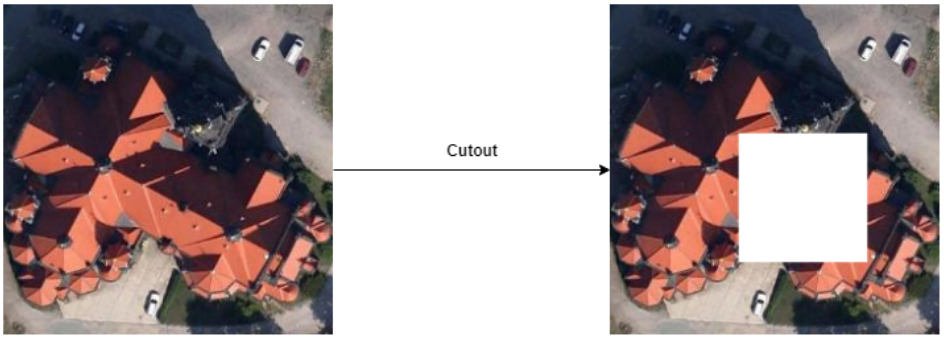


Figure 2.6: Cutout visualisation with remote sensing imagery.

2.5.2 Random Erasing

Random erasing was developed to enhance the existing cutout augmentation method. While cutout only focuses on removing randomly cropped areas of an image, random erasing focuses on a similar task but with the addition of noise/additional information to the cropped areas. The position and size of the area are also randomly chosen. This method also acts as a form of regularisation by introducing noise into the input images, which enables improved generalisation and performance gains [Zhong *et al.* 2020]. The random erasing method is visualised in figure 2.7.



Figure 2.7: Random Erasing visualisation with remote sensing imagery.

2.5.3 Mixup

The Mixup method utilises two input images from the training data and blends both together. This process generates a novel image for training [Zhang *et al.* 2017]. The novel training samples are generated as follows:

$$\begin{aligned} \tilde{x} &= \lambda x_i + (1-\lambda)x_j, \text{ where } x_i, x_j \text{ represent raw input vectors.} \\ \tilde{y} &= \lambda y_i + (1-\lambda)y_j, \text{ where } y_i, y_j \text{ represent one-hot label encodings.} \end{aligned} \tag{2.2}$$

where (x_i, y_i) and (x_j, y_j) are two training samples that are arbitrarily chosen and $\lambda \in [0, 1]$. Therefore, the Mixup method linearly mixes feature vectors and their labels. The method acts as a data-dependent form of regularisation by smoothing the decision boundary between classes, and makes the network robust to variations. It is a procedure that is simple to execute and is not computationally intensive. The Mixup method is visualised in figure 2.8.

2.5.4 CutMix

CutMix is a powerful data augmentation method that mixes the input images and labels. Entire regions within an input image are replaced and the label is adjusted in proportion to the area occupied by each class. This method allows a network to learn generalised features that are robust to noise and less prone to overfitting [Yun *et al.* 2019]. In the first row of Figure 2.9, the primary image features a church, partially substituted with a segment of a bird image. Similarly, the label of the final image is adjusted to reflect the ratio of both images. Additionally, the work of Uddin *et al.* [2020] improves upon CutMix by strategically utilising a saliency map to locate a representative image patch to be pasted on the primary image. This method leads to better feature learning.

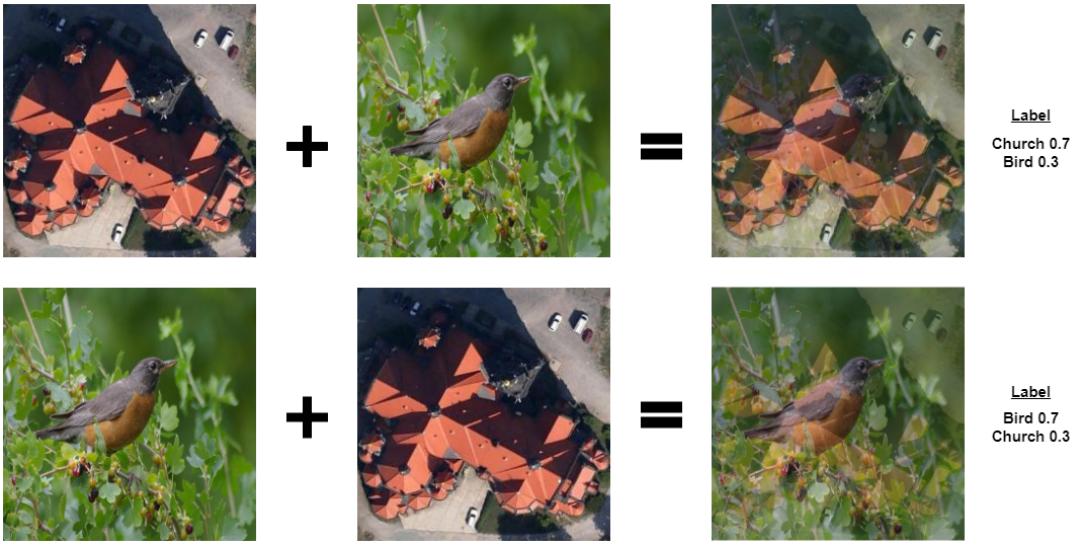


Figure 2.8: MixUp visualisation with remote sensing and natural imagery.

Another interesting paper that utilises ideas of Yun *et al.* [2019] is the work of Nakamura *et al.* [2022] which proposes a data augmentation method called Cross-Domain CutMix. This method aims to solve a limited labelled dataset and knowledge transfer problem. It combines CDFS learning with the CutMix data augmentation method to create a cross-domain augmentation method. A limited amount of images are incorporated into the source dataset from a target dataset during the training process, this enables patches of images from one domain to be pasted randomly onto the images from another domain during each training iteration. The cut region from one image is randomly selected and then positioned carefully onto the other image to ensure the important objects are not completely obscured. Network effectiveness is then determined with unseen classes from the target dataset. This method leads to enhanced feature representations being learnt and better adaptability when exposed to new classes from the target domain.

The data augmentation methods discussed can be implemented easily, add to the labelled data made available for training and improve network performance. Therefore, one component of the research will focus on improving the cross-domain data augmentation work seen in [Nakamura *et al.* 2022] and CutMix [Yun *et al.* 2019] to help solve the CDFS learning problem. An in-depth description of the enhanced method is provided in section 3.6.2.

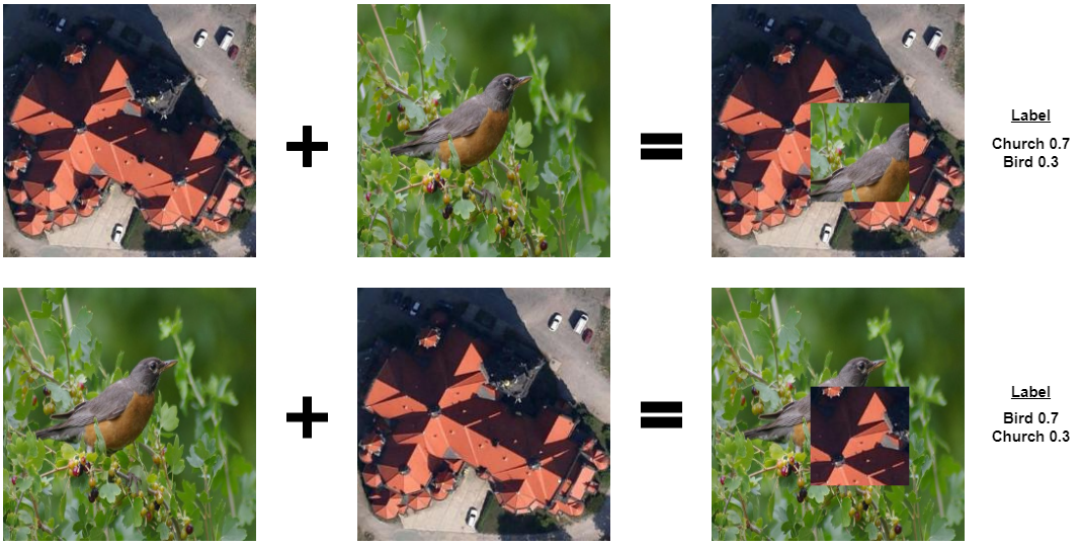


Figure 2.9: CutMix visualisation with remote sensing and natural imagery.

2.6 Summary

There is an abundance of existing research available for supervised and few-shot approaches that perform classification tasks on remote sensing imagery, with a number of innovative unsupervised and generative approaches that have been released within the past few years. Since the research in the CDFS learning field is still maturing, a limited amount of research exists for remote sensing scene classification. Therefore, this research will contribute to the growing research field of cross-domain and few-shot related topics within computer vision.

Chapter 3

Methodology

3.1 Introduction

This chapter will describe the methodology used to implement the proposed method. Firstly, the research hypothesis will be discussed, with a preview of the proposed method and its components. Secondly, relevant datasets to be used in our research, their descriptions and pre-processing methods are discussed. Thirdly, all of the components that will be applied in the proposed method for CDFS scene classification will be formally introduced and discussed in detail. Lastly, the experimental setup, assumptions, limitations and scope of the research will be discussed.

3.2 Research Hypothesis

The main aim of this dissertation is the validation of the following hypothesis:

> A CNN enhanced with advanced data augmentation and attention techniques can effectively address significant domain shifts and achieve comparable or superior performance to existing well-known methods in cross-domain few-shot scene classification tasks with a limited amount of labelled remote sensing data.

Validating the hypothesis necessitates several modifications to traditional CNN architectures and remote sensing image classification datasets. Specifically:

- Cross-Domain Region MixUp (CDRM) - A related cross-domain augmentation method by Nakamura *et al.* [2022] is introduced in section 2.5.4, but issues exist with the current method. Firstly, regions are randomly chosen to be cut from a target domain sample and pasted on a sample from the source domain. Secondly, pasting of the cutout region from the target domain sample may obscure

important features when pasted on the source domain sample. These issues are addressed in section 3.6.2.

- Dynamically Focused Attention Module (DFAM) - The attention modules introduced by Woo *et al.* [2018] in section 2.3.2 utilise all of the channel information regardless of importance and only consider spatial information at a single scale. These issues are addressed in section 3.6.3.
- Adaptive Task Aware Loss (ATAL) - Cross-entropy loss is effective (loss functions are discussed in section 2.3.4) but it does not account for individual task complexity during training and is only highly effective at inter-class separation. These issues are addressed in section 3.6.4.

3.3 Data

There are four datasets that will be used in this dissertation:

- Aerial Image Dataset (AID) [Xia *et al.* 2017].
- Remote Sensing Image Scene Classification dataset by Northwestern Polytechnical University (NWPU-RESISC45) [Cheng *et al.* 2017].
- Remote Sensing Image Classification Benchmark (RSI-CB) [Li *et al.* 2017b].
- Mini-Imagenet [Vinyals *et al.* 2016].

The first three datasets consist of remote sensing images sorted into classes for the purpose of classification tasks, each dataset will act as a target dataset during testing. The fourth dataset consists of natural images which will be used as our base/source dataset when performing CDFS learning.

3.4 Data Description

3.4.1 Aerial Image Dataset (AID)

The AID is classified as a large-scale aerial image dataset, with an RGB format that was constructed using Google Earth imagery [Xia *et al.* 2017]. The dataset consists of 10000 images labelled into 30 classes and represents a variety of aerial scene categories. All imagery in that dataset was labelled by experts from the remote sensing domain. AID has a multi-resolution characteristic, the images range from 8 metres to 0.5 metres per pixel in terms of spatial resolution, while the actual size of images in AID are 600 $\times$ 600 pixels.

3.4.2 Remote Sensing Image Scene Classification dataset by Northwestern Polytechnical University (NWPU-RESISC45)

The NWPU-RESISC45 is an extensive dataset comprising remote sensing scene images in a RGB format. The dataset was created by remote sensing specialists using Google Earth and has a total of 31500 labelled images that are categorised into 45 classes . Every class comprises 700 images, each with a dimension of 256 $\times$ 256 pixels. Images in this dataset also have a spatial resolution that varies from 30 metres to 0.2 metres per pixel for a majority of the classes [Cheng *et al.* 2017].

3.4.3 Remote Sensing Image Classification Benchmark (RSI-CB)

The RSI-CB is an extensive benchmark dataset which consists of remote sensing scene data in a RGB format and has been designed for classification tasks within the remote sensing field. The dataset was consolidated using multiple remote sensing platforms that account for a variety of geographical and environmental conditions. The dataset has a total of 24000 labelled images that are split into 35 classes with a resolution of 256 $\times$ 256 pixels [Li *et al.* 2017b].

3.4.4 Mini-ImageNet

The Mini-ImageNet is a popular dataset that is utilised to benchmark classification related algorithms. The dataset is a subset of the original ImageNet dataset, which consists of a 1000 classes and over a million labelled images [Deng *et al.* 2009]. Mini-ImageNet consists of 60000 RGB images split into 100 classes with 600 images per class and a resolution of 84 $\times$ 84 pixels [Vinyals *et al.* 2016]. Therefore, it is more manageable for training and evaluation of few-shot learning based networks.

3.5 Data Pre-processing

The source dataset (Mini-ImageNet) and target datasets (AID, NWPU-RESISC45 and RSI-CB) are split into test, validation and training sets. The test and validation sets will consist of 20% of the total classes while the training set will consist of images from the remaining classes. The dataset splits can be found in appendix A. Images from all sets will be re-scaled to 224 $\times$ 224 pixels before they become input for the network. All datasets are normalised and training images also undergo regular augmentation before they are used for the training process. This is in line with the same pre-processing method from Chen *et al.* [2019] used for their few-shot learning network. The regular augmentation process aims to increase the diversity and amount of data we have for training in addition to advanced augmentation techniques that will be applied, this is in line with the setup and process as seen in augmentation research (section 2.5).

3.6 Proposed Network

This section discusses the proposed network. Firstly, a preliminary discussion distinguishes CDFS learning from related learning approaches. Secondly, the outline of the proposed network utilised to solve the research problem is introduced, followed by an in-depth discussion of each addition that will be applied in the proposed network to validate the hypothesis stated in section 3.2.

3.6.1 Network Outline

The network proposed to solve the research problem will consist of three major components that are integrated into a CNN. The first component is the most important and consists of the CDFS learning method combined with a data augmentation method (CDRM), which is aimed at reducing domain shift by improving the adaptation ability of the network. The second component involves an improved feature extraction network, which consists of a ResNet18 backbone structure [He *et al.* 2016] in addition to improved convolutional block attention modules (DFAM) to increase the discriminative power of the overall network. The third component involves an adaptive task aware loss (ATAL) to improve the adaptive abilities of the network when it encounters new tasks and novel classes. Overall, the proposed method uses a prototypical structure [Snell *et al.* 2017] as a base with the additional components described above. The outline of the proposed network structure is visualised in Figure 3.1.

3.6.2 Cross-Domain Region MixUp

The method presented below is an advanced domain-level augmentation method that allows the CNN to learn more robust and effective feature representations. It aims to explicitly decrease the shift/gap between two data domains at the image level, which is caused by the differences in the feature distributions of the source and target domain datasets. Consequently, the method facilitates the network's effective adaptation to target domains, which is one of the main challenges when solving a CDFS learning problem. This method was independently developed but shares a degree of similarity with Nakamura *et al.* [2022] and improves on their work. Their focus is on object detection, and an object-aware cross-domain CutMix data augmentation technique is one of their proposed methods to address domain adaptation challenges and improve generalisation.

The augmentation method introduced in Nakamura *et al.* [2022] is based on the original CutMix method, as described in section 2.5.4. It is intended to mix features at an image level rather than the feature level since it operates on the input images. This process helps a model learn more robust invariant features and improves its overall gen-

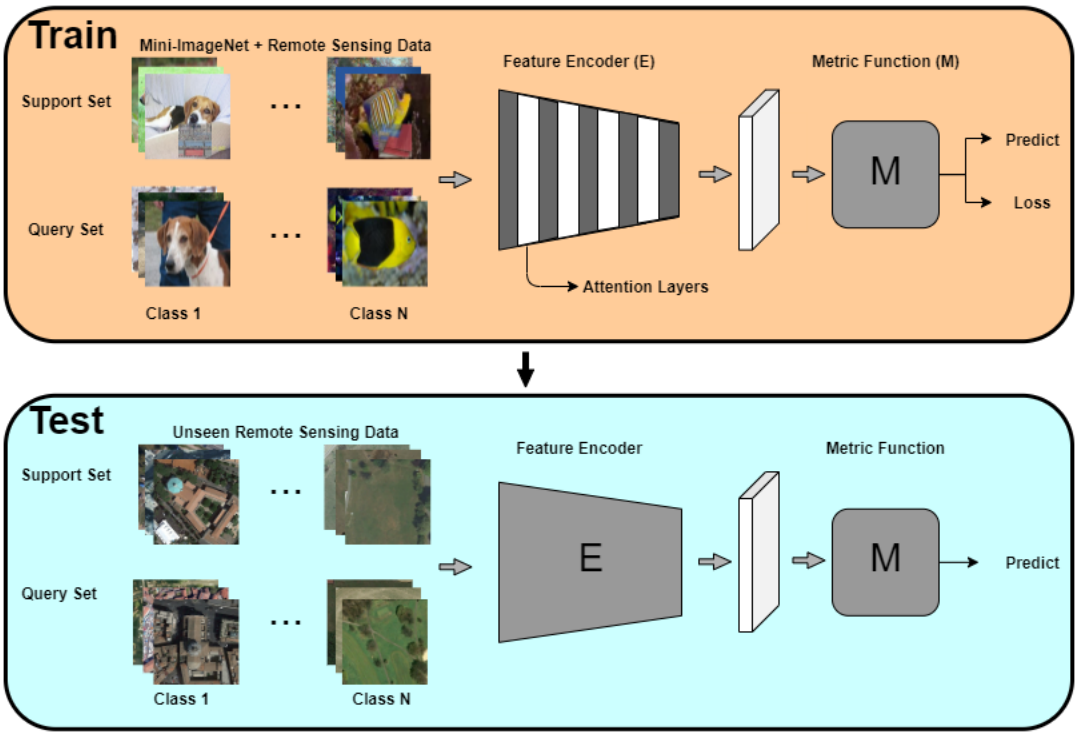


Figure 3.1: The diagram depicts the structure of the input data and the CNN, which emphasises the training and testing procedure used to solve CDFS classification.

eralisation ability during testing. Nakamura *et al.* [2022] extends the original CutMix to a cross-domain learning scenario to effectively handle domain adaptation challenges by mixing features at the domain level. This was achieved through the selective cutting of regions from a few images within a target domain and pasting them onto images from the source domain during each training iteration.

There are a range of issues that were identified with the methodology of Nakamura *et al.* [2022]. Firstly, the region of the image that was cut from the target domain was random and may not contain any important features. Secondly, when the cut region had to be pasted onto an image from the source domain, it could be placed directly over existing objects/features and this would impact model performance. To try and overcome this issue, they re-selected pasting positions by reducing the percentage of the objects within the original image being obscured but it is not optimal given the extra computational cost to perform the operation. Thirdly, the pasted region of the image still causes occlusion of objects regardless of its position within the source domain image due to the size and nature of the pasted region. The proposed Cross-Domain Region Mixup (CDRM) method attempts to solve these issues while improving model performance.

CDRM involves 5 main steps when the CNN is trained. Consider image A, that originates from a remote sensing target domain and an image B that belongs to the natural image source domain to be augmented during a training iteration for each batch:

1. λ is randomly drawn from a Beta distribution. It represents the proportion of mixing that will occur between image A and B.
2. A saliency map is then generated to identify the most salient region in image A, and a bounding box is produced with an area proportional to λ around the salient region.
3. This highly salient region from image A is then cut and pasted onto image B.
4. The cut and pasted region from image A is then blended with image B using λ. Blending (creating linear combinations) only occurs in the pasted region within image B.
5. The value of λ is then re-calculated to reflect the true proportion of the mixed area relative to the entire image. This is to ensure that the labels and the mixed image correspond with the actual mixing ratio.

The process described above is visualised in figure 3.2. The first row represents a process that occurs using an input image from a target domain. It depicts a saliency map generated from image A with the bounding box around the highly salient region that will be cut and pasted onto the input image from the source domain. The second row depicts the process that occurs with the source domain image, it shows how the selected region from image A is pasted onto image B and then blended.

CDRM addresses all of the issues that occur in the existing method. The first issue is addressed through the use of a saliency-guided bounding box which selects the most relevant regions from the target domain images during training, which improves the diversity of features that are learnt by the CNN. The second and third issues are solved through the blending step of the pasted region with the image from the source domain, which effectively prevents occlusion and eliminates the need for re-positioning of the pasted region. The mixing of features from different domains directly within training tasks, at the image level using CDRM, aims to help further improve the network's ability to adapt. It also ultimately leads to meaningful domain invariant features being learnt that enhances the network's generalisation ability during testing.

There are two disadvantages associated with CDRM. Firstly, the calculation of saliency maps and the blending step add to the computational overhead of the CNN during training. Secondly, the extra steps involved lead to increased complexity. Although the performance gain of this method outweighs the disadvantages, the effectiveness of CDRM on the CNN is emphasised in the ablation study (section 4.8.1).

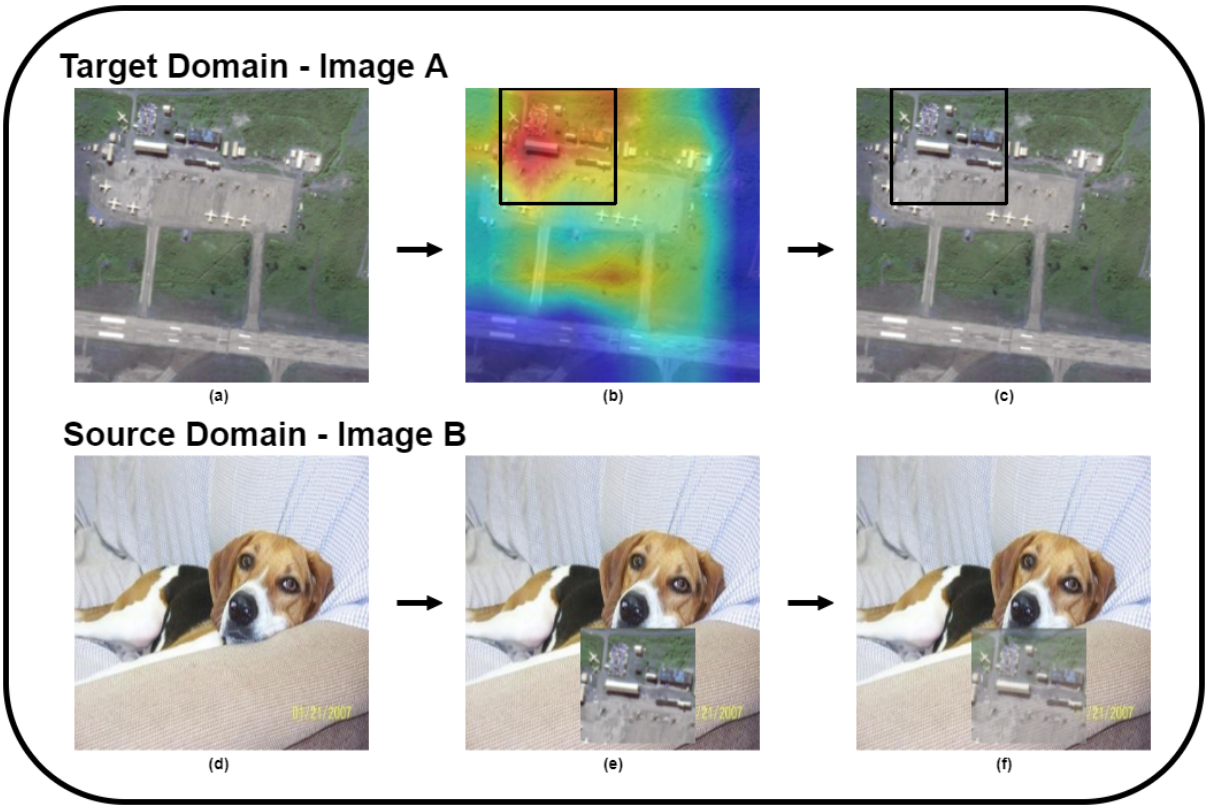


Figure 3.2: This diagram depicts the process on input images using CDRM. The first row represents the process that the target domain input image undergoes, while the second row represents changes that occur with respect to the source domain input image.

3.6.3 Dynamically Focused Attention Module

The ResNet18 backbone is the choice of CNN in this research, which will be fitted with the attention mechanisms that will be proposed in this section. Firstly, the network can be defined as G_θ with learnable parameters θ, for feature extraction of the remote sensing image input for the network. The addition of attention mechanisms, as discussed in section 2.3.2, improves CNN classification abilities. They have also been proven effective in increasing discriminative power of networks, which assists in alleviating issues associated with remote sensing imagery such as inter-class similarity and intra-class variance. Methods introduced in Woo *et al.* [2018] were selected to build upon and harness the properties of attention mechanisms. They will then be applied within a CNN to perform CDFS scene classification tasks.

Several issues were identified with the attention mechanisms introduced by Woo *et al.* [2018] in terms of channel and spatial feature processing. Firstly, a static approach was used for channel selection within the channel attention module. The approach does not make provision for considering input complexity, which is the varying characteristics of feature maps, nor does it achieve optimal efficiency by focusing on allocation of computational resources based on feature channel importance. Secondly, the spatial attention module uses a single-level approach and will only capture spatial features at

a single scale and one level of detail. This may cause the CNN to inadequately capture spatial patterns and dependencies. Therefore, in this section, the Dynamically Focused Attention Module (DFAM) is the mechanism proposed to solve these issues.

DFAM takes a feature map defined as $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ as input and then derives two feature maps, which will be discussed in the following sections. The first is a 1-dimensional channel attention mapping defined as $\mathbf{M_{channel}} \in \mathbb{R}^{C\times 1\times 1}$ and the second is a 2-dimensional spatial attention map defined as $\mathbf{M_{spatial}} \in \mathbb{R}^{1\times H\times W}$. This entire procedure is defined below:

$$\begin{aligned} \mathbf{F}' &= \mathbf{M_{channel}}(\mathbf{F}) \otimes \mathbf{F}, \\ \mathbf{F}'' &= \mathbf{M_{spatial}}(\mathbf{F}') \otimes \mathbf{F}', \\ \mathbf{F}''' &= \mathbf{F}'' + \mathbf{F}, \end{aligned} \tag{3.1}$$

where the $\otimes$ represents element-wise multiplication. In the multiplication process, channel attention values are transmitted via the spatial dimension, while spatial attention values are transmitted through the channel dimension. After the application of these processes, the output $\mathbf{F}''$ is obtained and combined with the feature map $\mathbf{F}$ to produce $\mathbf{F}'''$, which is the final output. This final step represents a residual connection to further improve feature learning through refinements learnt by previous layers [Liu *et al.* 2020], which makes the attention mechanism more robust. The structure of DFAM is visualised in figure 3.3.

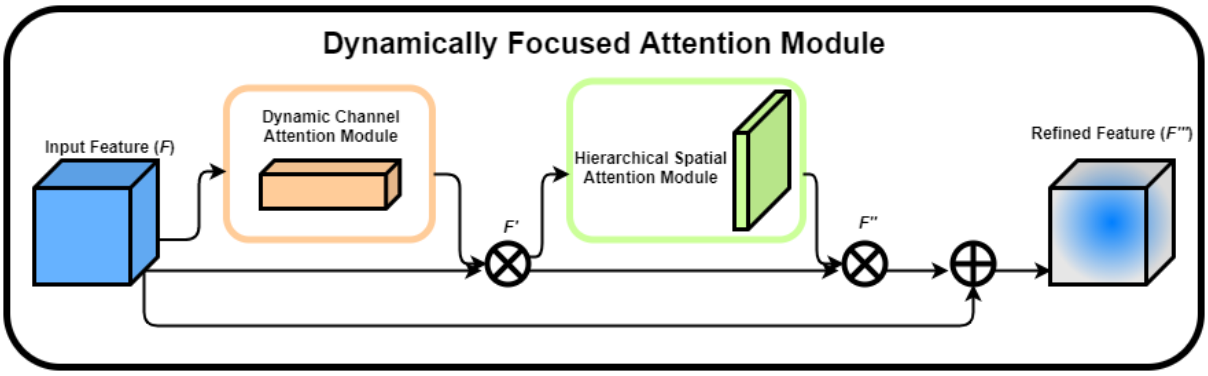


Figure 3.3: The DFAM contains the dynamic channel attention module and hierarchical spatial attention module. The diagram depicts the process where DFAM takes an input feature and produces the refined feature output.

3.6.3.1 Dynamic Channel Attention

The first component of DFAM is a dynamic channel attention module (DCAM) that consists of a channel attention module and a channel selector. It produces a channel attention map as output, which helps determine what features are present while empha-

sising the relevant channels according to the input complexity. This assists in improving the network's overall representational capacity. The channel attention module takes a feature map $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$ as input then derives two 1-dimensional channel attention maps, $\mathbf{M_c} \in \mathbb{R}^{C \times 1 \times 1}$ and $\mathbf{M_d} \in \mathbb{R}^{C \times 1 \times 1}$ as visualised in figure 3.4. This procedure is defined below:

$$\mathbf{M_{channel}}(\mathbf{F}) = \mathbf{M_c} \otimes \mathbf{M_d}, \tag{3.2}$$

To compute the first channel map, $\mathbf{M_c}$, the spatial dimension of the input feature map is compressed. The first step involves the simultaneous use of adaptive average-pooling and max-pooling, which enables effective feature aggregation [Woo *et al.* 2018]. This leads to the creation of two distinct spatial feature descriptors, which are defined as $\mathbf{F^c_{max}}$ for max-pooling and $\mathbf{F^c_{avg}}$ for average-pooling. A multi-layer perceptron (MLP) is employed in the next step. It has a single hidden layer with the layer size fixed at $\mathbb{R}^{C/r \times 1 \times 1}$, where $\mathbf{r}$ represents a reduction ratio that reduces parameter quantity. Note that the linear layers are used in the MLP. Both feature descriptors act as input for the MLP to generate a channel attention map $\mathbf{M_c} \in \mathbb{R}^{C \times 1 \times 1}$. The output feature vectors are subsequently combined using element-wise summation. This channel attention procedure is defined below:

$$\begin{aligned} \mathbf{M_c}(\mathbf{F}) &= \sigma\left(MLP\left(MaxPool(\mathbf{F})\right) + MLP\left(AvgPool(\mathbf{F})\right)\right) \\ &= \sigma\left(\mathbf{W_1}\left(\mathbf{W_0}\left(\mathbf{F^c_{max}}\right)\right) + \mathbf{W_1}\left(\mathbf{W_0}\left(\mathbf{F^c_{avg}}\right)\right)\right), \end{aligned} \tag{3.3}$$

where σ represents a sigmoid function, $\mathbf{W}_0 \in \mathbb{R}^{C/r \times C}$, and $\mathbf{W}_1 \in \mathbb{R}^{C \times C/r}$. Note that MLP weights are shared for both inputs since they are passed through the same network.

The second channel map $\mathbf{M_d}$ acts as a mask and only focuses on the most relevant channels. The number of channels to select is defined by $\mathbf{N_{scores}}$. It is computed by first setting the proportion of most relevant channels that will be kept, this variable is denoted by α_1 and is set to a value of 0.9 so that not more than 10% of channels are dropped. Therefore $\mathbf{N_{scores}}$ is the product of α_1 and the maximum number of initial feature map channels. This procedure is defined below:

$$\mathbf{N_{scores}} = int(\alpha_1 \cdot max_channels) \tag{3.4}$$

where the int function truncates the value towards zero. The next steps involve the use of the channel selector module (CSM) to calculate the importance of each channel within an initial feature map ($\mathbf{F}$). The first step involves the use of adaptive average-pooling for spatial aggregation and reducing a feature map along the spatial dimension, this results in the creation of spatial feature descriptor $\mathbf{F^c_{avg}}$. The CSM has a single hidden layer with the layer size fixed at $\mathbb{R}^{C/r \times 1 \times 1}$, where $\mathbf{r}$ represents a reduction ratio.

Note that layers in the CSM are convolutional in nature. The feature descriptor acts as input for the CSM to produce channel selection scores $\mathbf{C_{scores}} \in \mathbb{R}^{C\times1\times1}$. This procedure is defined below:

$$\begin{aligned}\mathbf{C_{scores}} &= \sigma(CSM(AvgPool(\mathbf{F}))) \\ &= \sigma(\mathbf{W_3}(\mathbf{W_2}(\mathbf{F^c_{avg}})))\end{aligned} \tag{3.5}$$

where σ represents the sigmoid function, $\mathbf{W}_2 \in \mathbb{R}^{C/r\times C}$ and $\mathbf{W}_3 \in \mathbb{R}^{C\times C/r}$. The indices with the most important channels $\mathbf{C_{selected}}$ are chosen using the top $\mathbf{N_{scores}}$ channels from the channel selection scores $\mathbf{C_{scores}}$. In other words, only the $\mathbf{N_{scores}}$ most important channels will be chosen. The mask $\mathbf{M_d} \in \mathbb{R}^{C\times1\times1}$ is initialised with zeros and then using $\mathbf{C_{selected}}$, the relevant indices are populated with ones. The mask $\mathbf{M_d}$ is then applied to the feature map $\mathbf{M_c}$ using element-wise multiplication, this results in the creation of feature map $\mathbf{M_{channel}}$ as seen in equation (3.2). This process employs concepts from channel gating/pruning to prioritise the critical channels and ignore the least critical channels.

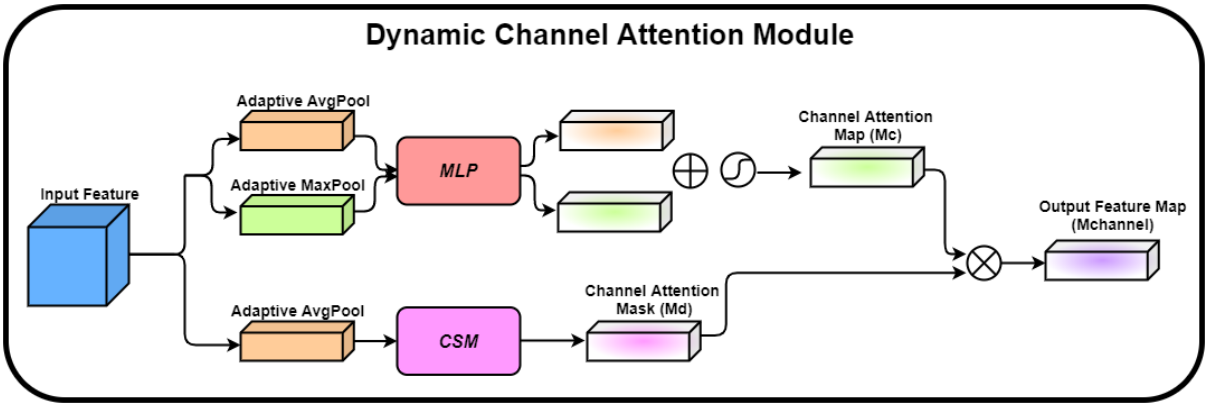


Figure 3.4: The diagram represents the Dynamic Channel Attention Module structure. It shows how the input feature is utilised to generate $\mathbf{M_c}$ in the top branch and $\mathbf{M_d}$ in the lower branch to produce the final output feature map.

3.6.3.2 Hierarchical Spatial Attention

The second component of DFAM is a hierarchical spatial attention module (HSAM), which prioritises where features are within an image and complements DCAM. The computation of the hierarchical spatial attention map $\mathbf{M_{spatial}}$ involves a feature map $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ passing through two levels within the module. Firstly, max-pooling and average-pooling are applied to reduce the feature map along the channel dimension and combine the output into a single feature descriptor. In order to emphasise features at varying scales, the descriptor is passed through two convolution layers with varying filter sizes to produce a final feature map $\mathbf{M_{spatial}}(\mathbf{F}) \in \mathbb{R}^{H\times W}$.

Firstly, channel information from a feature map is aggregated which generates a 2-dimensional map that represents features present in the different channels. The first is $\mathbf{F}^{\mathbf{s}}_{\mathbf{avg}} \in \mathbb{R}^{H \times W}$ the result of average-pooling and the second is $\mathbf{F}^{\mathbf{s}}_{\mathbf{max}} \in \mathbb{R}^{H \times W}$ which is the result of max-pooling. The two maps are then combined and convolved through the first convolutional layer to produce a 2-dimensional spatial map $\mathbf{M_{s1}}$. This output is then convolved through the second convolutional layer which uses a reduced filter size that produces another 2-dimensional spatial map $\mathbf{M_{s2}}$. Finally, the learned spatial maps are used to modulate the input feature map and produce the final output. This procedure is defined below:

$$\begin{aligned}\mathbf{M_{s1}}(\mathbf{F}) &= \sigma(f^{7\times7}([MaxPool(\mathbf{F}), AvgPool(\mathbf{F})])) \\ &= \sigma(f^{7\times7}([\mathbf{F}^{\mathbf{s}}_{\mathbf{max}}, \mathbf{F}^{\mathbf{s}}_{\mathbf{avg}}])),\end{aligned} \tag{3.6}$$

$$\mathbf{M_{s2}}(\mathbf{F}) = \sigma(f^{3\times3}(\mathbf{M_s}(\mathbf{F}))), \tag{3.7}$$

$$\mathbf{M_{spatial}}(\mathbf{F}) = \mathbf{F} \otimes \mathbf{M_{s1}}(\mathbf{F}) \otimes \mathbf{M_{s2}}(\mathbf{F}), \tag{3.8}$$

where σ represents a sigmoid function, $f^{7\times7}$ represents an initial convolutional operation that uses a filter size of 7×7 and $f^{3\times3}$ represents a second convolutional operation that uses a filter size of 3×3. The hierarchical spatial attention module is visualised in figure 3.5.

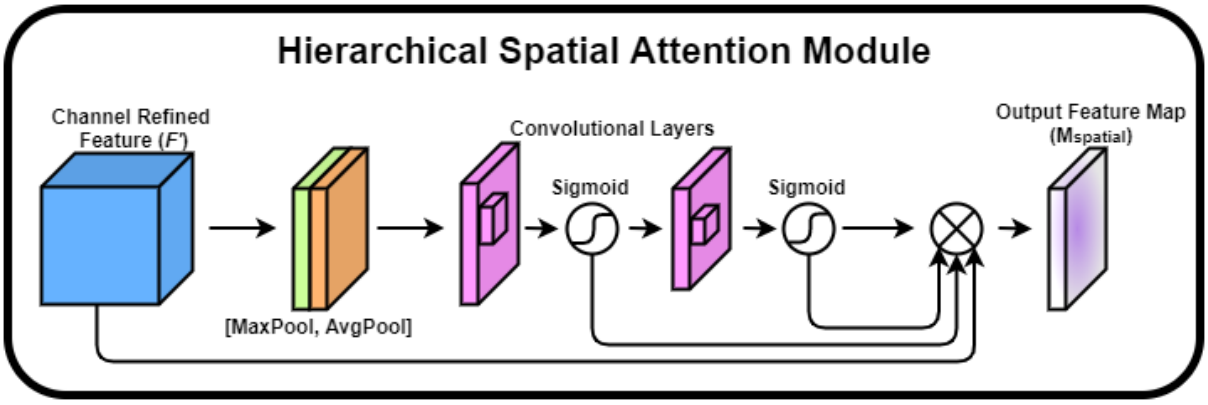


Figure 3.5: The diagram depicts the structure of the Hierarchical Spatial Attention Module. It shows how the channel refined feature is utilised to generate the output feature map $\mathbf{M_{spatial}}$.

DFAM addresses both of the existing issues with Woo *et al.* [2018], that were highlighted at the beginning of this section and are associated with their proposed module. The first issue is addressed through the application of the DCAM, where the chan-

nel selector module leverages the channel attention mechanism output to account for the input complexity of an image, which assists the network by focusing on the most relevant channels and reducing computational overhead. It is important to note that pooling within the channel attention modules and the dynamic channel selection serve different purposes. Pooling focuses on the reduction of spatial dimensions of a feature map to retain features and applies a static operation regardless of the input complexity. On the other hand, dynamic channel selection operates on the channel dimensions. It represents an adaptive operation and selects relevant channels based on the input complexity of a feature map.

The second issue is addressed by the HSAM that utilises a multi-level approach by using filters of different sizes within the module. It allows features of varying scales to be recognised, while ensuring spatial patterns and dependencies are effectively captured, which is essential given remote sensing image properties. One of the disadvantages of DFAM is that while the most relevant channels are dynamically selected, the top 90% of the most relevant channels are always selected. There may be input cases where more or less than 90% of the relevant channels can be selected to improve the network's discriminative power and efficiency.

Once the feature extraction is complete using the CNN with the improved attention mechanisms described above, the network will flatten the 512-dimensional embedding/output features and pass it onto the third component of the network.

3.6.4 Adaptive Loss Function

3.6.4.1 Metric Learning Classifier

The feature embedding/output $\mathbf{G}_\theta(\mathbf{x})$ serves as input for the metric learning classifier. The classifier, with the chosen metric being the Cosine similarity, measures the similarity between the feature embeddings of the support set and query samples. This process determines how close or similar a query sample is to each support class in the embedding space.

The process of using the metric learning classifier will be described in detail within this section. Firstly, a subset $\mathbf{S_c}$ can be denoted for all samples with a specific class $\mathbf{c}$ from the support set $\mathbf{S}$. Therefore, a *prototype* $\mathbf{P_c}$, can be defined as the mean of the output features that belong to $\mathbf{S_c}$ from the feature extractor $\mathbf{G}_\theta$. $\mathbf{P_c}$ is therefore the centroid of the class $\mathbf{c}$ and can be expressed as:

$$\mathbf{P_c} = \frac{1}{|\mathbf{S_c}|} \sum_{x_i \in S_c} \mathbf{G}_\theta(\mathbf{x_i}), \tag{3.9}$$

where $|\mathbf{S_c}|$ represents the number of samples in subset $\mathbf{S_c}$ and $\mathbf{x_i}$ denotes a sample within that subset. The probability that a sample $\mathbf{x}$ from the query set $\mathbf{Q}$ belongs to a specific class $\mathbf{c}$ can then be calculated by comparing the distance between the centroid $\mathbf{P_c}$ of class $\mathbf{c}$ and the feature embedding $\mathbf{G}_\theta(\mathbf{x})$. For Cosine similarity the prediction probability that the true label $\mathbf{y}$ of sample $\mathbf{x}$ is $\mathbf{c}$ can be expressed as:

$$\rho(\mathbf{y} = \mathbf{c}|\mathbf{x}) = \frac{exp(CS(\mathbf{G}_\theta(\mathbf{x}), \mathbf{P_c}))}{\sum_{c'} exp(CS(\mathbf{G}_\theta(\mathbf{x}), \mathbf{P_{c'}}))}, \tag{3.10}$$

where $CS(.,.)$ represents the Cosine similarity and $\mathbf{c}'$ denotes the index over all the classes. Given the explanation above, during each training task, the support set and the query set are constructed and passed to the ResNet18 backbone $\mathbf{G}_\theta$ for feature extraction. The feature embedding $\mathbf{G}_\theta(\mathbf{x})$ from $\mathbf{G}_\theta$ is then passed onto the metric learning classifier that will calculate the distance between a class centroid $\mathbf{P_c}$ and query embedding $\mathbf{G}_\theta(\mathbf{x_q})$ to predict labels for the query set. Therefore, to minimise the loss $\mathbf{L}$, it is calculated for each task and the parameters θ are updated through regular backpropagation and gradient descent algorithms. The next section describes the chosen loss function to be minimised during the training of the model.

3.6.4.2 Adaptive Task Aware Loss

The original cross-entropy loss has been proven effective in classification based tasks, it also has many advantages as discussed in section 2.3.4. Due to the loss function's characteristics, it has been chosen as the starting point to enhance network classification ability within the CDFS setting. The original cross-entropy function is described below:

$$\mathbf{L_{CE}} = -\frac{1}{N}\sum_{i=1}^{N} log(\rho(\mathbf{y_i}|\mathbf{x_i})), \tag{3.11}$$

where $\mathbf{N}$ is the number of query sample and $\rho(y_i|x_i)$ denotes the predicted probability that query sample $\mathbf{x_i}$ has the correct label $\mathbf{y_i}$.

There are a number of issues with the existing cross-entropy loss function that are present within a CDFS setting. The issues identified are as follows, firstly, the cross-entropy function processes each task and all samples similarly and does not account for individual task complexity during training. Therefore, this is an issue as the CDFS setting involves task-based training, and cross-entropy loss is less effective when feature distributions of the training and testing data are significantly dissimilar [Xu *et al.* 2021a]. This may reduce the generalisation ability of the network during testing. Secondly, the cross-entropy loss is effective at inter-class separation, but it is unable to

capture intra-class variance [Li and Wang 2020], which is a key characteristic of remote sensing imagery. Therefore, in this section Adaptive Task Aware Loss (ATAL) is the method proposed to solve these issues.

The improved ATAL loss function aims to improve upon the original cross-entropy loss by introducing information about the variability of tasks during training so that the network can improve its ability to handle CDFS learning scenarios. This is accomplished by extending the cross-entropy through the introduction of a regularisation term, $\mathbf{L_{intra}}$, that is adjusted based on the complexity of a task. The complexity, in this case, is determined by computing the mean intra-class variance within the support set. This process involves assessing feature distribution within each class and then taking the average across all the classes to obtain a adaptive loss term, $\mathbf{L_{intra}}$, for use in the loss function. The overall process can be represented as:

$$\mathbf{L_{ATAL}} = \mathbf{L_{CE}} + \alpha_2 \cdot \mathbf{L_{intra}}, \tag{3.12}$$

where $\mathbf{L_{ATAL}}$ is the Adaptive Task Aware Loss. It is comprised of $\mathbf{L_{CE}}$ which is the traditional cross-entropy as seen in equation (3.11), $\mathbf{L_{intra}}$ denotes the adaptive intra-class loss term based on the task complexity and $\alpha_2 \in [0, 1]$ is a scaling factor utilised to modulate the effect of $\mathbf{L_{intra}}$. The addition of an adaptive intra-class term adaptively scales the loss function based on the observed difficulty, which enables better feature representations to be learnt when there is high intra-class variance in a task. It balances the need for discriminative and compact feature representations, which leads to improved overall adaptability of the network.

The adaptive intra-class loss term is calculated in three steps using the support set that consists of 3 dimensions. This includes the number of classes, features and samples per class. The first step of the process involves the computation of the mean, $\mu_{\mathbf{ik}}$, of the *k*-th feature in the *i*-th class. The procedure is defined below:

$$\mu_{\mathbf{ik}} = \frac{1}{N} \sum_{j=1}^{N} \mathbf{x_{ijk}}, \tag{3.13}$$

where $\mathbf{x_{ijk}}$ represents the *k*-th feature of the *j*-th sample in the *i*-th class and $\mathbf{N}$ is the number of samples per class. The second step involves the computation of the variance, $\sigma^{\mathbf{2}}_{\mathbf{ik}}$, of the *k*-th feature in the *i*-th class. The procedure is defined below:

$$\sigma^{\mathbf{2}}_{\mathbf{ik}} = \frac{1}{N} \sum_{j=1}^{N} (\mathbf{x_{ijk}} - \mu_{\mathbf{ik}})^2, \tag{3.14}$$

The third and final step involves the computation of the mean across all classes and features, L_{intra}, using the output from the second step. This is necessary to collapse the intra-class variances into a single value that can be used within the loss function. The procedure is defined below:

$$L_{intra} = \frac{1}{C \cdot F} \sum_{i=1}^{C} \sum_{k=1}^{F} \sigma_{\mathbf{ik}}^{\mathbf{2}}, \tag{3.15}$$

where $\mathbf{C}$ is quantity of classes, $\mathbf{F}$ is number of features and $\sigma_{\mathbf{ik}}^{\mathbf{2}}$ is the intra-class variance.

The ATAL function is able to address the first issue, outlined above, where the initial cross-entropy loss ($\mathbf{L_{CE}}$) does not account for individual task complexity during training. ATAL accomplishes this through the introduction of an adaptive intra-class loss term ($\mathbf{L_{intra}}$) which complements the original cross-entropy term depending on task complexity. A larger adaptive intra-class loss term suggests a more complex task therefore, the overall loss will be increased to reflect this change per training task. The model will be able to adapt and focus on these complex tasks and improve how it learns from them. Therefore, the cross-entropy loss and adaptive intra-class loss term forces the network to improve how it learns the overall variability within tasks during training.

The second issue outlined above, of the original cross-entropy only being effective at inter-class separation, is addressed by the adaptive intra-class loss term itself as it contains a measure of the intra-class variance of the support set. A higher variance suggests that the spread of the examples within a class is large. This additional information in the loss function is essential to the model as it allows for improved and balanced learning to account for the feature distribution within the classes in each task. Overall, ATAL leads to a more robust and accurate classification of remote sensing imagery as it also helps address intra-class and inter-class characteristics of a dataset by leveraging task information.

3.7 Experimental Setup

The experimental setup is discussed in detail in this section. Firstly, the networks that will be compared to the proposed network are discussed. It includes a baseline, traditional and other related networks. Secondly, the implementation details are discussed in detail to ensure the reproducibility of results. Lastly, the evaluation details are discussed to outline which metrics will be used to measure performance and ensure fair comparisons between networks.

3.7.1 Baseline Network and Comparison Architectures

The CNN architectures that will be utilised in comparison to the proposed network have been designed to satisfy requirements of the experiments that are conducted in the dissertation. Architectures discussed in this chapter all feature a ResNet18 backbone for feature extraction due to the popularity and usage of shallow networks in existing research [Chen *et al.* 2019].

A baseline network is required to establish the actual effectiveness of the proposed network, it provides a reference point to measure improvements introduced by the proposed approach. The baseline network used in this dissertation is designed to take input that is passed onto the feature extractor which is a regular ResNet18 backbone. The feature embedding/output is then passed onto the fully connected layer followed by the softmax layer. This network is trained in a supervised manner using the source base class data $\mathbf{D}^{\mathbf{base}}_{\mathbf{Source}}$ through the minimisation of a regular cross-entropy loss. Testing is then conducted within the CDFS setting with target novel class data $\mathbf{D}^{\mathbf{novel}}_{\mathbf{Target}}$. The baseline network follows a similar setup to Chen *et al.* [2019] and will help confirm the efficacy of a CDFS learning approach that is proposed.

Alongside the baseline network there are three comparison architectures, which are traditional few-shot approaches. These include a prototypical network [Snell *et al.* 2017] with the Cosine similarity metric, a matching network [Vinyals *et al.* 2016] and a relation network [Sung *et al.* 2018] with softmax loss. These traditional architectures were chosen due to their popularity and effectiveness, and will be used to benchmark against the proposed method. Other comparison methods include three augmentation methods, MixUp [Zhang *et al.* 2017], CutMix [Yun *et al.* 2019], Cross-Domain CutMix [Nakamura *et al.* 2022] and an attention method that consists of CBAM [Woo *et al.* 2018]. These architectures are included due to their similarity with components of the proposed method.

The ProtoNet is the chosen base architecture for the augmentation and attention methods with the Cosine similarity as the distance function. This choice is due to the proposed network also using a ProtoNet architecture base as mentioned in section 3.6.1. The augmentation and attention architectures are crucial to validating the efficacy of the proposed network as it aims to improve on elements of related architectures/methods.

3.7.2 Implementation Details

The experimental setup used is similar to Chen *et al.* [2019], this includes the chosen quantity of epochs/tasks for the training phase, testing phase and learning rate etc. CDFS classification is performed on the three remote sensing datasets that act as target datasets (AID, NWPU-RESISC45 and RSI-CB) and source dataset (Mini-ImageNet). Training for the Baseline network takes place over 400 epochs for all N-way K-shot setting and batch size has a set value of 16. The meta-learning methods and pro-

posed network also have a batch size of 16, but are trained over 600 epochs (60000 tasks/episodes) for the 1-shot setting and 400 epochs (40000 tasks/episodes) for the 5-shot and 10-shot setting. It is important to note that 100 tasks are equivalent to 1 epoch. The validation set is employed during training to preserve the model version with the most optimal accuracy.

In each task, the N-way K-shot structure is followed, and this means that N classes are randomly sampled for an N-way classification task. N is set to 5 for all meta-training and meta-testing across a majority of experiments as this is a widely used setting across the few-shot learning literature, but the value of N can vary. Each of the classes will have K labelled instances per class as the support set where $K = \{1, 5, 10\}$. K is varied as the aim of the experiments is to gauge the performance of the architectures on the availability of labelled data. The query set will have 16 instances per K-shot task which is set to match the batch size, but it can be randomly chosen.

The Cross-Domain CutMix method [Nakamura *et al.* 2022] explores a variety of proportions of target data to be sampled and mixed with the source data during the training process. It is classified as single source where the target domain data is accessible [Wang *et al.* 2023a]. In this dissertation, all experiments will use a $\frac{1}{8}$ setting as the Cross-Domain CutMix proved to be effective with the limited amount of labelled data available. Therefore, the proposed network will also use the same setting in all of the experiments. This will demonstrate the efficacy of the proposed network using minimal data from a target dataset.

All of the methods in this paper are trained from scratch using a learning rate of 0.001. The proposed network will have the scaling factor α_2 fixed at a value of 1, the reasoning for this decision is explained in section 4.6. Additionally, an Adam optimiser is utilised across all methods. During meta-testing for all the methods, the N-way K-shot structure, support and query set are setup to match the meta-training, but it is sampled from randomly chosen classes in the novel test data. The meta-testing experiments are averaged over 600 tasks, and the results are reported as a mean accuracy with 95% confidence intervals. All of the code was implemented using the Python programming language with the Pytorch library and was run on a NVIDIA GeForce RTX 3090 GPU. For more information on the experimental setup and settings, refer to Chen *et al.* [2019].

3.7.3 Evaluation Details

For experiments the accuracy metric will be the main metric used for evaluation. However, the F1-score will also be considered to provide a holistic view of network performance. The accuracy metric is the most suitable evaluation metric for this research since it is based on classification. The F1-score is the harmonic mean of the precision and recall. Therefore, it ensures that the effect of false positives and false negatives are taken into account [Powers 2020].

Comparisons of results will be made between the proposed network and the architectures outlined in section 3.7.1 and experiments will be carried out using the different target datasets, with multiple N-way K-shot settings to be tested for all networks.

3.8 Research Limitations and Assumptions

This section elaborates on the research assumptions and limitations that define the scope of the dissertation. A comprehensive review of the research led to three main limitations that have been identified, which are as follows:

- Although CDRM, DFAM and ATAL could theoretically be adapted to various neural network architectures, this research is limited to applying CNNs to image classification. It is also crucial to acknowledge that efficacy of the additions could vary significantly when implemented in other settings.
- The focus of the dissertation is on CDFS classification, as well as improving the domain adaptation and generalisation ability of a network. Although, the ability of the proposed network may not translate to unseen domains (domains that are not present in this research). It is also important to note that this research focuses on solving a complex CDFS problem where there is a significant domain shift.
- The characteristics of target datasets and their relation to the source dataset in CDFS classification have an impact on efficacy. The distribution and size of a target dataset, inter-class similarities or intra-class variability can all have a major impact on the results that are achieved during testing.

Additionally, before experiments were conducted, a number of assumptions had to be established given the wide range of methods presented in modern research. These assumptions prevented the scope of the research from becoming unmanageable and increased the feasibility of running the experiments. According to a review of the literature discussed in chapter 2, the assumptions are as follows:

- It is assumed that the CDFS learning paradigm is suitable for scene classification in a limited label dataset domain such as remote sensing. This implies that few-shot learning can effectively learn to classify correctly from a reduced amount of data during the training phase. It is important to note that this research does not evaluate CDFS approaches in comparison to other contemporary approaches/paradigms.
- The second assumption is that the CDFS learning paradigm is suitable for handling real-world scenarios and reducing the effects of domain shift, which occurs due to differences between feature distributions of datasets. CDFS learning is able to simulate these scenarios well, with a source domain dataset used for training and a target domain dataset for testing.

- The third assumption is that CNNs are a reasonable starting point to introduce improved domain adaptation and generalisation within the CDFS learning paradigm. CNNs are popular due to their effective ability in the extraction of hierarchical features and can be expected to adapt and generalise well from a few labelled examples.

3.9 Summary

In this chapter, the research methodology and experimental setup were introduced, in addition to the remote sensing datasets that will be used in the experiments. Firstly, the research hypothesis was described together with the necessary modifications to achieve it. Secondly, the datasets and the relevant pre-processing steps are discussed. Thirdly, the outline of the network was clarified in and the three additions that will be applied to solve the research problem were discussed in detail. Fourthly, the experimental setup was discussed, which includes comparison architecture, implementation details and evaluation details. Lastly, the limitations, assumptions and scope of the dissertation were discussed.

Chapters 2 and 3 discussed the rationale behind the need for improvements and the descriptions of proposed solutions. In the following chapter, it will be empirically proved that the proposed solutions are effective as a whole, which will be followed by ablation studies to determine the additive effect of each solution. The experiments, which will be qualitative and quantitative, are going to be carried out on multiple datasets. This will ensure that any concerns regarding the functioning of the proposed solutions are alleviated.

Chapter 4

Results and Discussion

4.1 Introduction

In this section, the proposed network (CDAN), traditional meta-learning models, relevant data augmentation and attention methods are set up for CDFS scene classification and then tested in experiments. The first experiments, where access to labelled samples is varied, will feature each of the target datasets (AID, NWPU-RESISC45 and RSI-CB) and are the main experiments of this dissertation. Additional experiments that follow will focus on varying numbers of classes, the scaling factor for the loss term and network depth. Lastly, a comprehensive ablation study will be presented for CDAN. For each experiment the results will be presented and then discussed in detail.

4.2 AID

4.2.1 Results and Discussion

In this section the source dataset is Mini-ImageNet, while the target dataset is AID. Therefore, the architectures are trained in a CDFS classification setting and tested on novel/unseen classes from AID to produce the results in this section. The first CDFS experiment showcases the classification accuracy of the different architectures with confidence intervals across multiple 5-Way K-shot settings in Table 4.1. The second CDFS experiment showcases the F1-scores of the different architectures across multiple 5-Way K-shot settings in figure 4.1.

Table 4.1: CDFS classification accuracy test results on different architectures for AID with 5-Way K-Shot settings.

Architecture	**1-Shot**	**5-Shot**	**10-Shot**
Baseline	44.25% ± 0.71%	65.08% ± 0.60%	69.22% ± 0.50%
MatchingNet	47.09% ± 0.69%	65.57% ± 0.56%	72.40% ± 0.49%
RelationNet	42.06% ± 0.72%	65.61% ± 0.54%	70.54% ± 0.49%
ProtoNet	48.52% ± 0.64%	69.68% ± 0.57%	77.64% ± 0.45%
CBAM	44.97% ± 0.68%	70.22% ± 0.57%	76.52% ± 0.45%
MixUp	43.70% ± 0.64%	71.45% ± 0.52%	76.99% ± 0.45%
CutMix	43.91% ± 0.65%	70.99% ± 0.49%	77.80% ± 0.43%
Cross-Domain CutMix	51.56% ± 0.76%	76.96% ± 0.50%	80.95% ± 0.42%
CDAN (Ours)	**55.30% ± 0.76%**	**81.31% ± 0.41%**	**85.04% ± 0.34%**

The results from the first experiment (table 4.1) indicate that the classification accuracy improves as the number of shots also increases. It is observed that the Baseline architecture has an accuracy of 44.25%, 65.08% and 69.22% for the 1-shot, 5-shot and 10-shot settings respectively. This is a general trend that is observed across all architectures in the results, and it means all architectures perform better as the number of samples per class made available increases. The increase in available labelled data causes the architectures to capture more patterns and underlying features, which leads to improved feature representations that increases classification performance.

The Baseline architecture has the worst performance across most K-shot settings on AID compared to other architectures. This indicates the lack of suitability of supervised learning in a cross-domain setting for the identification of novel classes from AID. This finding is unsurprising, as supervised learning requires large labelled datasets to perform well, which was discussed in section 2.4.2.

The traditional architectures for few-shot learning, such as MatchingNet, RelationNet and ProtoNet, improve upon the baseline with the exception of RelationNet, which underperforms on the 1-shot (42.06%) setting. The best performing traditional architecture is ProtoNet across all K-shot settings (48.52%, 69.68%, 77.64%), but it still underperforms when compared to the CDAN. The results also show that the traditional architecture for few-shot learning is not as effective in cross-domain settings with AID, but the ProtoNet and MatchingNet architecture is still better than the supervised Baseline method. The lack of performance seen in the traditional few-shot architecture is most likely caused by a large domain shift that exists between the natural image source dataset (Mini-ImageNet) and the remote sensing target dataset (AID). The traditional methods are unable to effectively minimise the discrepancy between the two dataset domains. It implies that a more adaptable method is required to solve the problem.

CBAM, MixUp and CutMix have mixed results in comparison to the regular ProtoNet architecture. They all underperform in comparison to the regular ProtoNet architecture with the 1-shot setting. CBAM and MixUp only perform better using the 5-shot setting with 70.22% and 71.45% respectively, in comparison to the 69.68% achieved

by ProtoNet. They both underperform when the 1-shot and 10-shot settings are used in comparison to ProtoNet. CutMix is able to perform better than ProtoNet using the 5-shot and 10-shot settings (70.99% and 77.80%), but the improvement is marginal. These mixed results are interesting as the CBAM, MixUp and CutMix all use a prototypical base network, and they do not appear robust when compared to ProtoNet, which does not have any additions to improve performance.

The main cause of the mixed results with CBAM, MixUp and CutMix in table 4.1 is also because of the cross-domain setting and large domain shift that exists between the source and target dataset. The feature representations learnt are inadequate due to the architectures only being exposed to the source data, and may be overfitting on the source domain. Therefore, it shows that the benefits of the original attention and augmentation methods are not able to be effectively leveraged in a CDFS setting during testing with the novel classes from AID.

The first architecture to outperform ProtoNet and the original CutMix in all K-shot settings is Cross-Domain CutMix (51.56%, 76.96%, 80.95%) by over 3%. This performance is due to the advanced augmentation technique that makes it more robust to domain shift. Furthermore, CDAN is able to achieve superior performance across all K-shot settings (55.30%, 81.31%, 85.04%) when compared to all architectures. The increase in performance of CDAN can be attributed to the combined effect of the additions. It outperforms the Baseline network by over 10% and the Cross-Domain CutMix by over 3% for each K-shot setting.

The improved performance of Cross-Domain CutMix is due to the improved adaptability of the architecture within a cross-domain setting, which samples a set amount of images per class from the target data training set that leads to improved features being learnt as mentioned in section 2.5.4. However, the performance improvement of CDAN in the CDFS setting over all the architectures implies that it is even more effective when a large domain shift is present between the source domain dataset and the target domain dataset. In addition to this finding, the results clearly demonstrate that the additions are effective and CDAN is proficient with a limited quantity of labelled data.

The results for the second experiment in this section (figure 4.1) represents F1-score test results. They indicate a similar trend as seen in the classification accuracy results for AID. All of the architectures display an increase in F1 scores as the number of shots increases and more labelled data is available. The baseline once again has the lowest results in terms of F1 scores when compared to the other models, which implies it is not as effective in identifying actual positive cases while minimising the false positives as well as the false negatives. While the ProtoNet architecture is shown to be the best among the traditional methods it also performs better than CBAM, MixUp and CutMix in the 1-shot setting. Mixed results are seen among the regular attention and augmentation methods when compared to ProtoNet across the 5-shot and 10-shot settings.

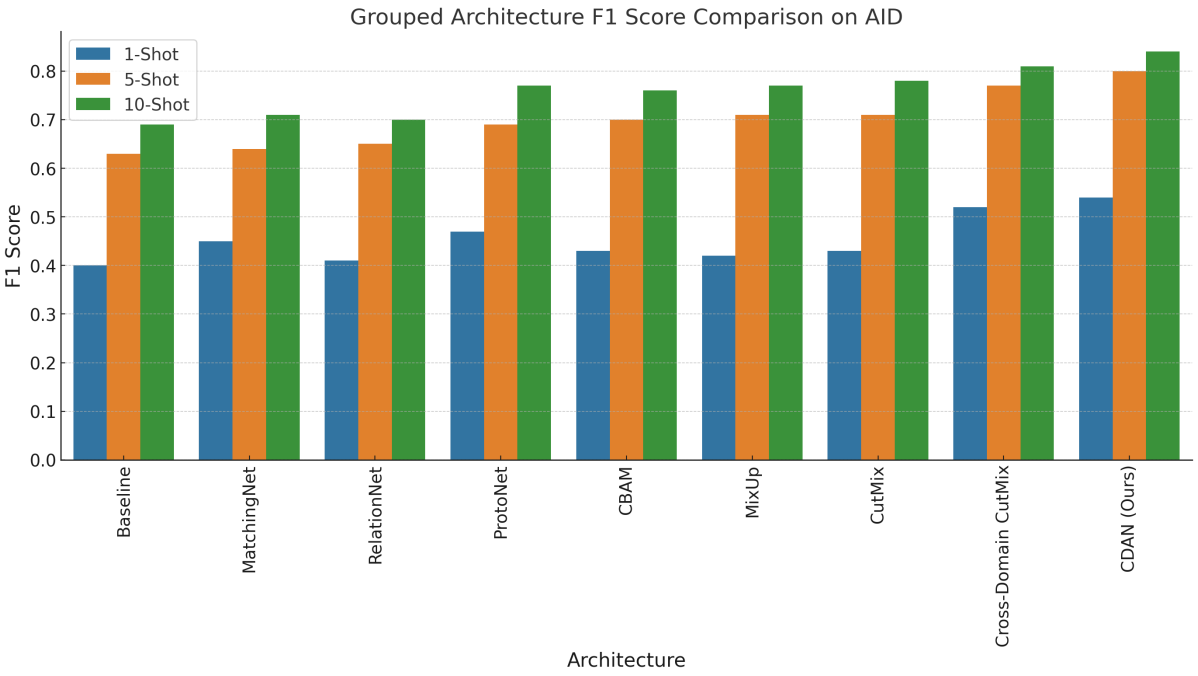


Figure 4.1: CDFS classification F1-score test results on different architectures for AID with 5-Way K-Shot settings, where K={1, 5, 10}

.

Cross-Domain CutMix has a better performance when compared to the regular augmentation and attention architectures across all shot settings. It is also the first existing architecture to achieve an F1 score over 0.8 for the 10-shot setting which indicates that it has decent performance. Although, it is clear that CDAN achieves the best F1 score with every K-shot setting when compared to the other architectures. While Cross-Domain CutMix showcases good performance when compared to the regular augmentation and attention architectures across all shot settings, the performance of CDAN indicates that it is a more robust network which is able to correctly predict positive cases.

The main aim of the experiments and results in this section was to determine the classification accuracy and ability of the complete CDAN in comparison to traditional few-shot architecture and related methods. The proposed network has demonstrated that it is able to overcome domain adaptation challenges and showcase improved generalisation ability during testing on the unseen classes of the AID, even with a limited amount of labelled data.

4.3 NWPU-RESISC45

4.3.1 Results and Discussion

In this section the source dataset is Mini-ImageNet, while the target dataset is NWPU-RESISC45. Therefore, the architectures are trained in a CDFS classification setting and tested on novel/unseen classes from NWPU-RESISC45 to produce the results in this section. The first CDFS experiment shows the classification accuracy of the different architectures with confidence intervals across 5-Way K-shot settings in table 4.2. The second CDFS experiment depicts the F1-scores of the different architectures for multiple 5-Way K-shot settings as seen in figure 4.2.

Table 4.2: CDFS classification accuracy test results on different architectures for NWPU-RESISC45 with 5-Way K-Shot settings.

Architecture	**1-Shot**	**5-Shot**	**10-Shot**
Baseline	40.64% ± 0.74%	60.70% ± 0.79%	72.09% ± 0.70%
MatchingNet	47.84% ± 0.89%	65.00% ± 0.75%	67.50% ± 0.76%
RelationNet	40.48% ± 0.79%	63.75% ± 0.78%	66.09% ± 0.79%
ProtoNet	44.40% ± 0.76%	65.11% ± 0.78%	73.96% ± 0.68%
CBAM	44.95% ± 0.78%	64.12% ± 0.80%	70.46% ± 0.75%
MixUp	43.83% ± 0.78%	66.91% ± 0.79%	74.14% ± 0.73%
CutMix	45.54% ± 0.82%	68.77% ± 0.72%	71.60% ± 0.69%
Cross-Domain CutMix	60.69% ± 0.94%	79.50% ± 0.66%	83.73% ± 0.56%
CDAN (Ours)	**62.06% ± 0.91%**	**83.56% ± 0.59%**	**86.63% ± 0.51%**

The results from the first experiment (table 4.2) indicate that the classification accuracy improves as the number of shots increases. It is observed that MatchingNet has a classification accuracy of 47.84%, 65% and 67.50% for the 1-shot, 5-shot and 10-shot settings respectively. This is a general trend that is observed across all architectures in results and is because of increases in available samples per class for different K-shot settings. This trend is due to the improved feature representations that are learnt as more samples per class are used, which leads to improved overall classification performance.

The first experiment also shows that the Baseline architecture has the worst performance with the 5-shot setting on NWPU-RESISC45 compared to other architectures. Surprisingly, the Baseline architecture performs better than MatchingNet on the 10-shot setting. Additionally, it outperforms RelationNet in both the 1-shot and 10-shot settings. The Baseline architecture has the worst overall performance on NWPU-RESISC45 compared to CDAN, even though it has a slightly better performance on certain K-shot settings when compared to some traditional architectures. This finding is consistent with observations from the literature in section 2.4.2 and the results in section 4.2, as supervised methods require large amounts of labelled data to perform well.

MatchingNet and ProtoNet outperform the Baseline on the 1-shot and 5-shot setting, while the best performing traditional architecture overall is the ProtoNet with 65.11% for 5-shot and 73.96% for the 10-shot setting. The results from table 4.2 also show that the traditional methods for few-shot learning are not that effective in a CDFS setting with NWPU-RESISC45. The reason for the lack of performance is due to the large domain shift that exists between the source dataset (Mini-ImageNet) and the target dataset (NWPU-RESISC45). Therefore, the traditional methods are unable to effectively minimise the discrepancy between the two domains due to their designs. This necessitates the need for adaptability of the architecture to effectively CDFS classification.

The CBAM, MixUp and CutMix architectures have mixed results when compared to the ProtoNet and only completely outperform the Baseline on the 1-shot and 5-shot settings. CBAM only performs better in the 1-shot setting with 44.95% compared to ProtoNet with 44.40%. MixUp performs better in both the 5-shot and 10-shot settings with 66.91% and 74.14% respectively, in comparison to the 65.11% and 73.96% achieved by ProtoNet. Lastly, the CutMix architecture performs better than ProtoNet in the 1-shot and 5-shot settings with 45.54% and 68.77%, respectively. These mixed results are interesting as the CBAM, MixUp and CutMix all use a prototypical base network, which indicates that the regular attention and augmentation methods are not as effective as expected, when compared to a regular Protonet architecture without any additions.

The leading cause of these mixed results for CBAM, MixUp and CutMix when compared to the ProtoNet is because of the cross-domain setting and significant domain shift that exists between the source and target dataset. As discussed in section 4.2.1, where this is also observed, the feature representations learnt are inadequate when exposed to novel remote sensing data. This is due to CBAM, MixUp and CutMix only being exposed to the source data, and may be overfitting on the source domain. Therefore, it is shown that the benefits of the original attention and augmentation methods are not able to be effectively leveraged in a CDFS setting during testing with the novel classes from NWPU-RESISC45.

Cross-Domain CutMix performs better than a majority of the architectures in table 4.2. It outperforms the baseline and traditional architectures discussed above, by more than 10% across all K-shot settings. It also outperforms related augmentation methods, MixUp and CutMix, by more than 10% in the 1-shot and 5-shot setting and by 9% for the 10-shot setting, which is significant. This is due to the improved augmentation ability of Cross-Domain CutMix, which makes it more adaptive to the target domain and unseen classes. However, CDAN is able to achieve superior performance across all K-shot settings (62.06%, 83.56%, 86.63%) when compared to all of the architectures due to the effect of the additions. It is able to gain more than 20% over the baseline for 1-shot and 5-shot sample options and over 10% for all other comparison architectures across all K-shot settings, except Cross-Domain CutMix, which is closer to its performance but still inferior.

The improved performance of Cross-Domain CutMix over previously discussed architectures in table 4.2 is due to the improved adaptability of the method within a cross-domain setting to the target domain. However, the superior architecture in this experiment is CDAN, due to the combined effects of the additions that work together to improve the architecture's adaptability when dealing with large domain shifts that exist between the training (source domain) and testing (target domain) data. CDAN is also more robust in comparison to existing architecture and related architecture, which implies that it can overcome limited labelled data environments such as the remote sensing field.

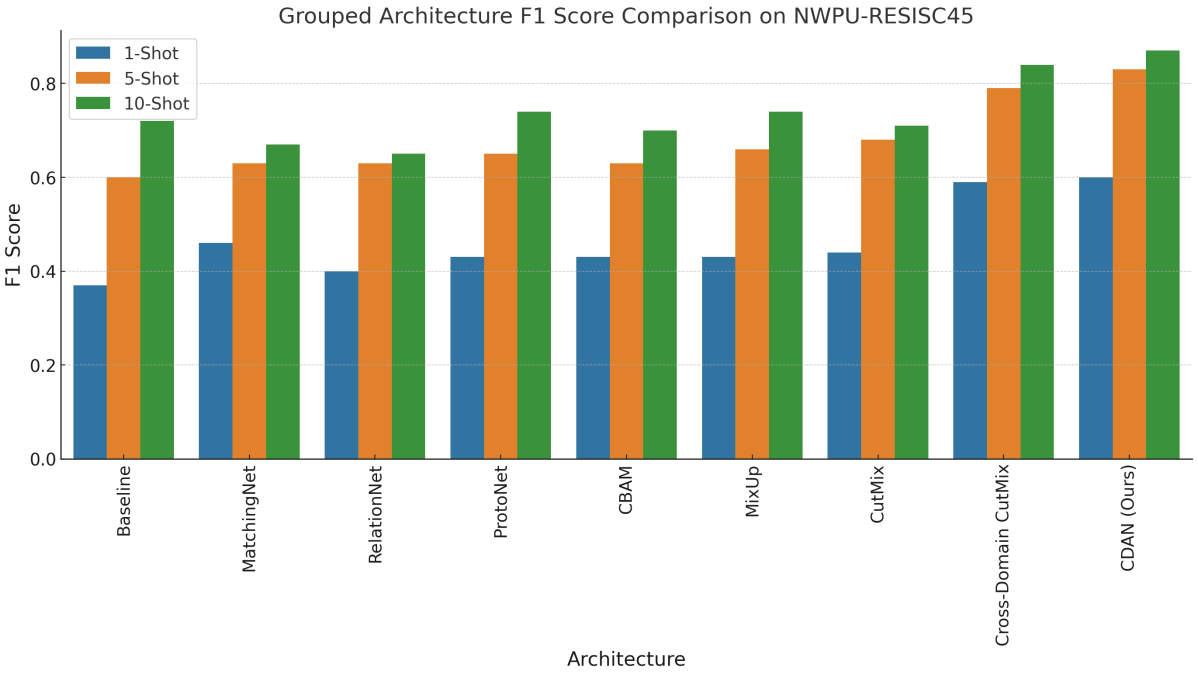


Figure 4.2: CDFS classification F1-score test results on different architectures for NWPU with 5-Way K-Shot settings, where K={1, 5, 10}

.

The results for the second experiment (figure 4.2) and represent the F1-score test results for NWPU-RESISC45. The results have a similar trend to the classification accuracy results. An increase in F1-score is observed for all of the architectures as the number of shots increases. The baseline has the lowest F1 scores for the 1-shot and 5-shot setting when compared to the other architectures but outperforms two of the traditional architectures (MatchingNet and RelationNet) as well as CBAM and CutMix in the 10-shot setting. This implies that the baseline is better than the architectures mentioned above at identifying positive cases while minimising false cases when more labelled data is available.

The ProtoNet is the only traditional architecture to perform better than the baseline across all K-shot settings. It also achieves higher F1 scores than CBAM, MixUp and CutMix in certain cases. The Cross-Domain CutMix is the first method to achieve higher

F1 scores than all of the previously discussed architectures. This can also be observed in figure 4.2, but ultimately, CDAN has the highest F1 scores when compared to all the architectures across all K-shot settings. Therefore, Cross-Domain CutMix and CDAN are the most balanced models when identifying correct positive cases during classification when compared to all the other architectures, which reaffirms the findings from the NWPU-RESISC45 classification accuracy results.

The experiments and results presented in this section help determine the classification accuracy and efficacy of CDAN in comparison to traditional as well as related architecture on the unseen classes of the NWPU-RESISC45 dataset. In these experiments, the proposed network has demonstrated the ability to overcome domain adaptation challenges and showcased improved generalisation ability during testing on the unseen classes of NWPU-RESISC45, with a limited amount of labelled data.

4.4 RSI-CB

4.4.1 Results and Discussion

In this section the source dataset is Mini-ImageNet, while the target dataset is RSI-CB. Therefore, the architectures are trained in a CDFS classification setting and tested on novel/unseen classes from RSI-CB to produce the results in this section. The first CDFS experiment in this section shows the classification accuracy of the different architectures with confidence intervals across 5-Way K-shot settings in table 4.3. The second CDFS experiment in this section depicts the F1-scores of the different architectures for multiple 5-Way K-shot settings as seen in figure 4.3.

Table 4.3: CDFS classification accuracy test results on different architectures for RSI-CB with 5-Way K-Shot settings.

Architecture	**1-Shot**	**5-Shot**	**10-Shot**
Baseline	67.32% ± 0.83%	87.75% ± 0.49%	91.46% ± 0.38%
MatchingNet	67.27% ± 0.74%	86.74% ± 0.51%	90.23% ± 0.39%
RelationNet	54.56% ± 0.78%	82.76% ± 0.50%	89.84% ± 0.42%
ProtoNet	67.74% ± 0.69%	90.87% ± 0.36%	92.02% ± 0.33%
CBAM	67.56% ± 0.76%	88.06% ± 0.45%	92.26% ± 0.32%
MixUp	65.98% ± 0.76%	86.19% ± 0.47%	91.71% ± 0.31%
CutMix	68.30% ± 0.77%	91.96% ± 0.36%	91.98% ± 0.31%
Cross-Domain CutMix	73.18% ± 0.73%	91.37% ± 0.36%	94.66% ± 0.28%
CDAN (Ours)	**76.03% ± 0.71%**	**93.18% ± 0.34%**	**95.51% ± 0.23%**

The results from the first experiment (table 4.3) indicates that the classification accuracy improves as the number of shots increases. It is observed that MatchingNet has a classification accuracy of 67.27%, 86.74% and 90.23% for the 1-shot, 5-shot and 10-shot settings, respectively. This is a general trend that is observed across all architectures in the results and is due to an increase in available samples per class for different K-shot settings. All architectures perform better with more data because the quality of the learnt feature representations improves, but the performance of all architectures with RSI-CB implies that the characteristics of this specific dataset have made it easier for them to adapt. This has led to accuracies of over 85% in the 5-shot and 10-setting on RSI-CB, which was only seen with CDAN in section 4.2 and section 4.3.

In the first experiment, the Baseline architecture (67.32%, 87.75%, 91.46%) outperforms MatchingNet and RelationNet across all K-shot settings. This finding is unexpected given the performance of the baseline in section 4.2 and section 4.3, which showed that it underperformed in most cases when compared to the traditional methods. Another observation is that ProtoNet (67.74%, 90.87%, 92.02%) is able to outperform the baseline, MatchingNet and RelationNet architectures across all of the K-shot settings. The results in table 4.3 show that the traditional methods are effective in a cross-domain setting with RSI-CB, given the over 90% accuracy achieved by most of the architectures in the 10-shot. This outcome is most likely due to the characteristics of the RSI-CB dataset, as mentioned above, which results in a smaller difference between feature distributions of the source domain dataset and the target domain dataset. The results show that the traditional methods are able to minimise the discrepancy between the two domains more effectively on this specific target dataset, which leads to improved performance. While the traditional architectures were able to achieve better results for CDFS classification, they are still inferior to CDAN.

CBAM and CutMix perform better when compared to the baseline across all K-shot settings but have mixed results when compared to ProtoNet. CBAM performs better with 92.26% in the 10-shot setting in comparison to ProtoNet with 92.02%, whereas CutMix performs better in the 1-shot and 5-shot setting with 68.30% and 91.96% respectively, in comparison to 67.74% and 90.87% achieved by ProtoNet. On the other hand, MixUp underperforms when compared to the baseline with the 1-shot setting and 5-shot setting by more than 1%, and it also underperforms across all K-shot settings when compared to ProtoNet. These mixed results indicate that the regular attention and augmentation methods are ineffective in a CDFS setting with remote sensing data such as RSI-CB, when compared to existing traditional methods like ProtoNet. These results are interesting as the CBAM, MixUp and CutMix all use a prototypical base network, which indicates that the regular Protonet architecture without any additions showcases better or similar performance.

The main cause of these mixed results for CBAM, MixUp and CutMix when compared to the Baseline and ProtoNet is because of the domain shift that exists between the source and target dataset, even if the domain shift is not as significant with RSI-CB. As discussed in section 4.2.1 and section 4.3.1, the feature representations learnt are inadequate when exposed to novel remote sensing data. This is due to the CBAM, MixUp

and CutMix only being exposed to the source data and may be overfitting on the source domain. While CBAM, MixUp and CutMix achieve relatively good performance during testing, it is shown that the benefits of these original attention and augmentation methods are not able to be effectively leveraged in a cross-domain setting during testing with the novel classes from RSI-CB. There is observed degradation in performance when compared to the original ProtoNet, which has no attention or advanced augmentation methods.

Cross-Domain CutMix performs better than the baseline and traditional architectures across all K-shot settings. It also outperforms CBAM and MixUp across all settings but underperforms against the original CutMix with the 5-shot setting by more than 0.5%. The results indicate that an improved augmentation method that can handle CDFS scenarios is able to perform better on the RSI-CB dataset. However, CDAN is still able to outperform all of the architectures across all K-shot settings (76.03%, 93.18%, 95.51%). Given the results across all architectures in table 4.3, the domain shift does appear to be smaller between the target and source datasets due to characteristics of RSI-CB, but the performance gain of CDAN can still be attributed to the combined effects of the additions. These additions have a role in improving the architecture's adaptability and performance when dealing with the domain shift that exists between training (source domain) and testing (target domain) data. It is also able to operate well in comparison to existing architecture, which implies that it is also able to overcome limited labelled data settings such as the remote sensing field.

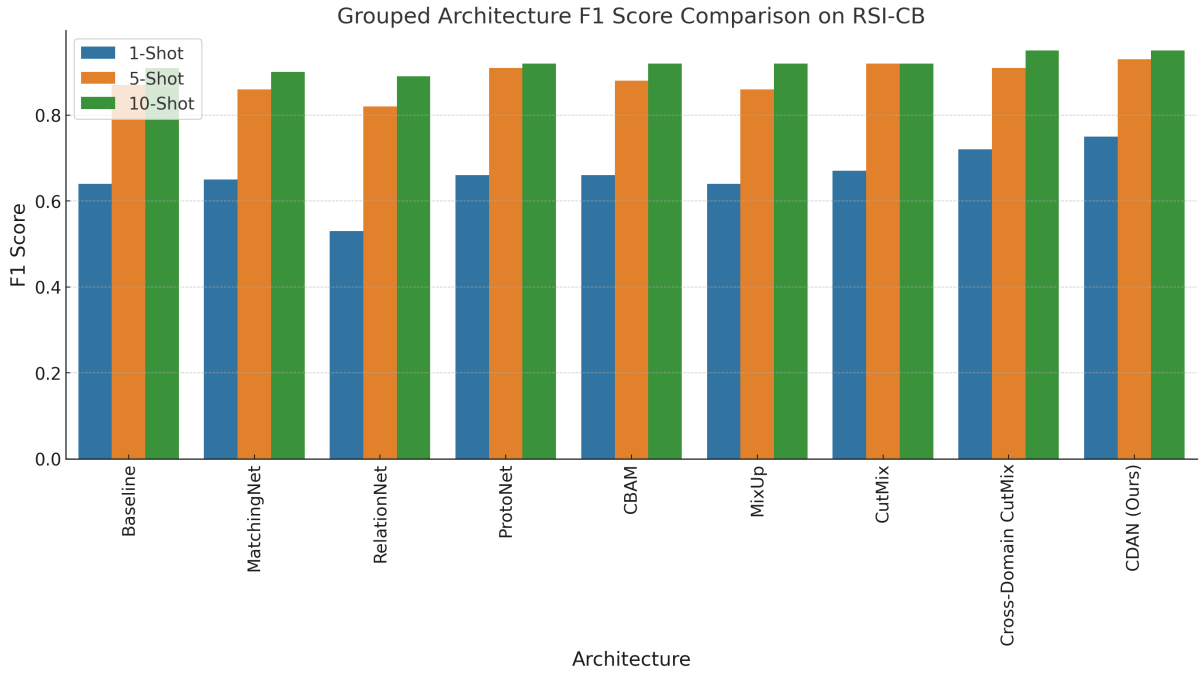


Figure 4.3: CDFS classification F1-score test results on different architectures for RSI-CB with 5-Way K-Shot settings, where K={1, 5, 10}

.

The results for the second experiment (figure 4.3) and represent the F1 score test results for RSI-CB. The results have a similar trend to the classification accuracy results.

An increase in F1 score is observed for all of the architectures as the number of shots increases. The improved performance is due to the increased number of labelled samples that are made available. The baseline has better F1 scores for the 5-shot and 10-shot settings when compared to MatchingNet and RelationNet and has similar scores to CBAM and MixUp across the K-shot settings. The high F1 scores for the baseline imply that it is good at identifying positive cases while minimising false cases when more labelled data is available for the RSI-CB dataset.

The ProtoNet is the only traditional architecture to perform better than the baseline across all K-shot settings. It is also able to achieve higher F1 scores than CBAM and MixUp in certain cases but performs worse when compared to CutMix. The Cross-Domain CutMix is the first method to achieve higher F1 scores than all of the previously discussed architectures, and this is due to its adaptation abilities in a cross-domain setting. However, CDAN has the highest F1 scores compared to all the architectures across all K-shot settings. Therefore, the proposed method is the best architecture for identifying correct positive cases during classification and reaffirms the findings from the RSI-CB classification accuracy results.

The main aim of these experiments and results is to determine the classification accuracy and ability of CDAN compared to traditional and related methods on the unseen classes of the RSI-CB dataset. In these experiments, CDAN has demonstrated that it is able to overcome domain adaptation challenges and showcased improved generalisation ability during testing on the unseen classes of RSI-CB, with a limited amount of labelled data.

It is also important to note that while attention modules were added to CDAN, which increased the number of learnable parameters, the increase in parameters is minimal compared to the other architectures. Therefore, the increased size is not responsible for the entirety of the performance gain seen with CDAN. The total number of learnable parameters can be found in table A.5, these parameters are applicable to all experiments in section 4.2, section 4.3, and section 4.4 for the respective architectures. All comparison architectures have the same number of learnable parameters with the exception of CBAM, which has a slightly higher number of parameters because it has attention modules. Therefore, the architecture is also useful when compared with CDAN.

The difference in the number of learnable parameters between CDAN and most of the comparison architectures is 1.58% while the difference is much smaller compared to CBAM at 0.79%. These differences are small in terms of architectures and can be compared effectively, given that they all have approximately 11 million learnable parameters. Furthermore, an extensive ablation study is carried out in section 4.8, which indicates that the slight increase in learnable parameters is not responsible for the performance gains associated with CDAN.

4.5 Effect of N-way Variations

The main experiments in the previous three sections focused on varying the number of labelled samples (K) while keeping the number of classes constant (N) during CDFS classification. The purpose was to gauge the performance of CDAN in relation to traditional and existing related methods in a challenging setting where labelled data was limited. However, in this section, the experiment carried out only involves the Baseline and CDAN while varying the number of classes (N) for the different K-shot settings. It seeks to highlight the performance of CDAN in relation to the Baseline as the number of classes are adjusted with different amounts of samples available during CDFS classification. Figure 4.4, figure 4.5 and figure 4.6 represent the results of the experiment for each target dataset (AID, NWPU-RESISC45 and RSI-CB). Note that the 1-way (single class) setting has been left out as it is a trivial case and will always be 100%.

Figure 4.4 represents the CDFS classification accuracy on AID with the Baseline and CDAN while varying N-way K-shot settings. The 5-way setting across the different K-shot settings represents the classification accuracy results of CDAN as shown in section 4.2. The first observation is that as the number of classes decrease from 5-way to 2-way there is an observed increase in classification accuracy across the 1-shot, 5-shot and 10-shot settings for the Baseline and CDAN. This is due to the classification task becoming less complex, the fewer the number of classes to predict makes it easier for the networks to accurately predict the correct class. However, as the number of classes escalates, classification tasks become increasingly more challenging, this leads to a reduction in accuracy. Another observation is that the Baseline performance decreases more than CDAN across all K-shot settings as the number of classes is varied.

Figure 4.5 and figure 4.6 represents the CDFS classification accuracy on NWPU-RESISC45 and RSI-CB with the Baseline and CDAN while varying N-Way K-Shot settings. The 5-way setting across the different K-shot settings for both figures represents the classification accuracy results of CDAN, as shown in section 4.3 and section 4.4. A similar observation to figure 4.4 is evident for figure 4.5 and figure 4.6. As the number of classes decreases, from 5-way to 2-way, there is a clear increase in classification accuracy across the 1-shot, 5-shot and 10-shot settings as the classification complexity of the tasks decreases. Inversely, when the numbers of classes escalates, classification tasks become increasingly more challenging, which leads to a decrease in classification accuracy. Additionally, it is once again observed that the Baseline performance is worse when compared to CDAN across all K-shot settings on both datasets as the number of classes is varied. This reinforces the finding, from previous sections, that a supervised training approach in a CDFS setting is inadequate.

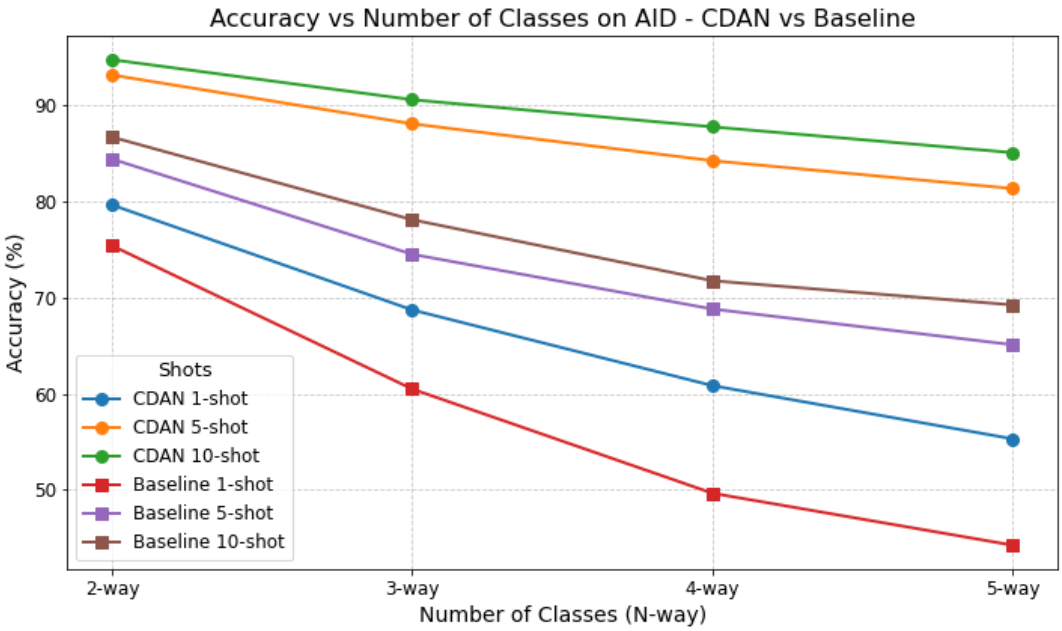


Figure 4.4: CDFS classification accuracy for the Baseline and CDAN on AID with N-way K-shot settings.

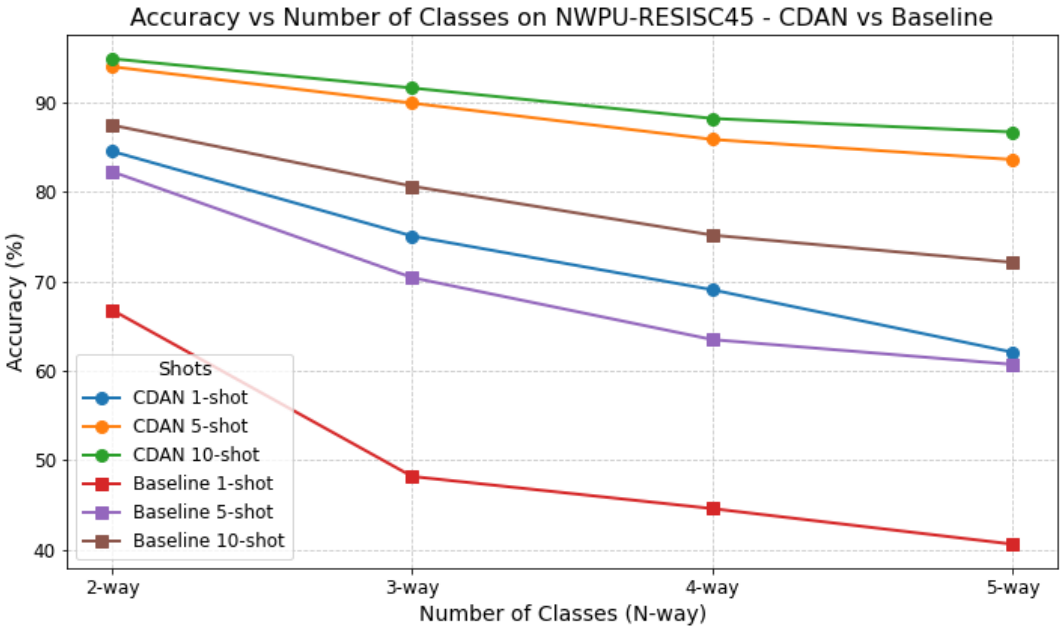


Figure 4.5: CDFS classification accuracy for the Baseline and CDAN on NWPU-RESISC45 with N-way K-shot settings.

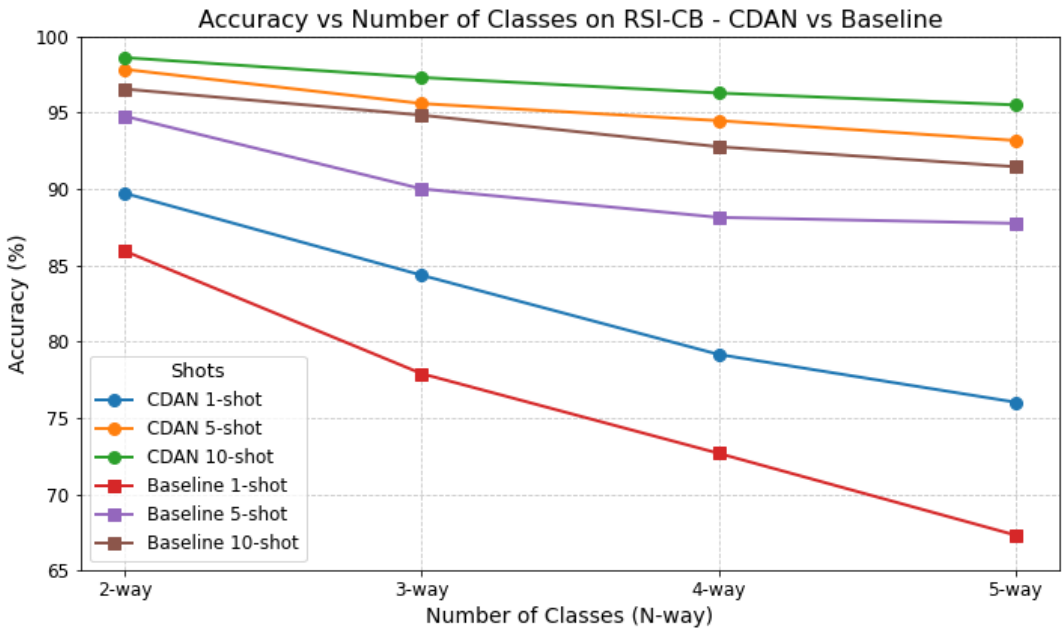


Figure 4.6: CDFS classification accuracy for the Baseline and CDAN on RSI-CB with N-way K-shot settings.

4.6 Effect of Scaling Factor α_2

This section presents an experiment that varies the scaling factor α_2. The scaling factor operates on the loss term, $\mathbf{L_{intra}}$ that was introduced in section 3.6.4, and influences how the network learns. The experiment is performed with CDAN using the 5-way 5-shot setting, which is only considered since it offers a balanced evaluation between the extremes of the 1-shot and the 10-shot setting. The 1-shot setting risks high variability since it only has one example per class, and the 10-shot setting might mask the effects of the scaling factor since it has a larger number of examples per class. Additionally, in practical CDFS scenarios, the five examples per class is a reasonable choice as one example may be too restrictive and ten examples may not always be feasible in a low data availability setting.

The experiment is performed using all three target datasets (AID, NWPU-RESISC45 and RSI-CB) to showcase the effect that the choice of scaling factor has on the classification performance of CDAN. Therefore, this experiment justifies the choice of α_2 that was utilised in the main experiments (section 4.2, section 4.3 and section 4.4), given the impact that it has on the accuracy and performance of CDAN.

Figure 4.7 represents the CDFS classification results of the experiment on CDAN using the 5-way 5-shot setting with the three target datasets. As the scaling factor is increased, for AID and NWPU-RESISC45, there is a clear fluctuation in classification accuracy. Ultimately, the scaling factor of 1 yields high accuracy on both datasets. This implies that during training CDAN benefits from the full effect of the loss term $\mathbf{L_{intra}}$.

Although CDAN has better performance on the RSI-CB dataset, the fluctuation in accuracy is also observed as the scaling factor is increased. The scaling factors with the highest accuracies for RSI-CB are similar. Therefore, a scaling factor of 1 would be the better option, given that it also yields the best performance on AID and NWPU-RESISC45. In conclusion, this experiment justified the choice of a scaling factor (α_2) set to 1 in all experiments that feature CDAN.

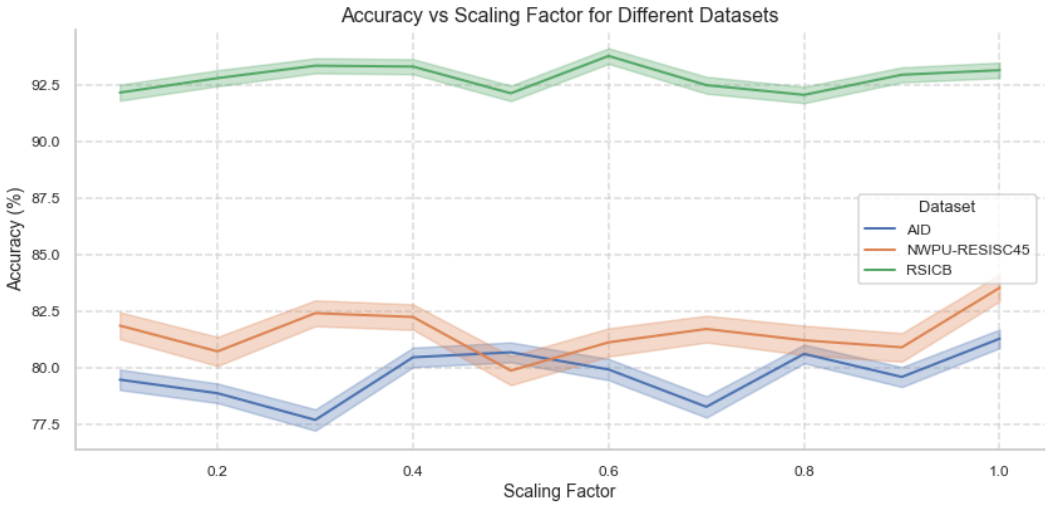


Figure 4.7: CDFS classification accuracy for all target datasets using CDAN with 5-way 5-shot settings while varying the scaling factor α_2.

4.7 Effect of Network Depth

In this section, a network depth experiment is performed with the proposed method (CDAN). The ResNet18 backbone is the chosen backbone in this dissertation as mentioned in section 3.7.1, but it is a shallow network. Therefore, the aim of this experiment is to determine the effect of larger backbones on the performance of CDAN in a CDFS classification setting. The experiment is run on all three target datasets (AID, NWPU-RESISC45 and RSI-CB) to provide a fair and comprehensive comparison.

Figure 4.8 represents the visual result of the CDFS classification experiment that was carried out with a 5-way 1-shot setting using the ResNet18, ResNet34 and ResNet50 backbones across all of the target datasets. The first observation with AID as the target dataset is that the accuracy decreases as the network backbone depth increases. Therefore, the ResNet18 achieves the best accuracy, while the ResNet50 achieves the lowest accuracy for the AID target dataset. The second observation from this experiment is that a similar trend is present with NWPU-RESISC45 as the target dataset. The ResNet18 achieves a better accuracy than the larger network backbones, and the ResNet50 has the worst performance. The third observation from this experiment is that while the results with the RSI-CB are higher than the other two target datasets, the trend seen across the different backbones is also present. Overall, the ResNet18 has a

higher accuracy when compared to the ResNet34 and ResNet50 backbones regardless of the target dataset with the 5-way 1-shot setting.

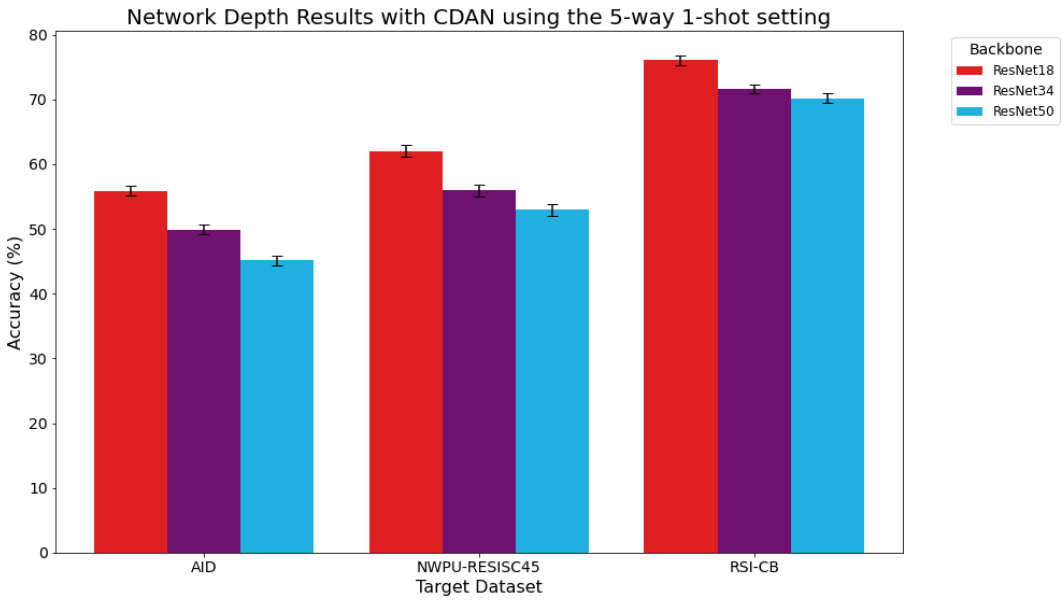


Figure 4.8: CDFS classification network depth test results on different target datasets using CDAN with 5-Way 1-Shot settings.

Similar trends are observed in figure 4.9 and figure 4.10 which depict the results for the network depth experiment using the 5-way 5-Shot and 5-way 10-Shot setting across the three target datasets. One interesting observation is the performance of the ResNet34 backbone on the RSI-CB dataset in figure 4.9, where the ResNet34 outperforms the ResNet18 and ResNet50 backbone. There are also smaller differences in accuracy as the network depth increases when more samples are made available, this can clearly be seen if figure 4.8 and figure 4.10 are compared. Overall, the conclusion is that an increase in ResNet network depth decreases the performance of CDAN. The reason for the better performance observed with the ResNet18 backbone may be due to their ability to adapt faster to a target domain encountered within a CDFS setting.

Deeper networks start to capture more specific and domain-dependent features in deeper layers, but this is not beneficial when encountering data from a different domain. The learnt domain-specific features may not be able to generalise well in deeper networks, leading to decreased performance in CDFS classification settings as shown in Cheng *et al.* [2021] with an increasing amount of ConvBlocks. Another reason for the performance of shallow networks, like the ResNet18, may be due to the rate at which they converge. Shallower networks generally converge quicker during training due to the smaller amount of parameters that need to be optimised, whereas deeper networks have a larger amount of parameters that need to be optimised and may take longer to converge. This characteristic of shallow networks may be an advantage in a CDFS setting, where there is a reduced amount of labelled data available for training, as solutions may be found faster than with deeper networks.

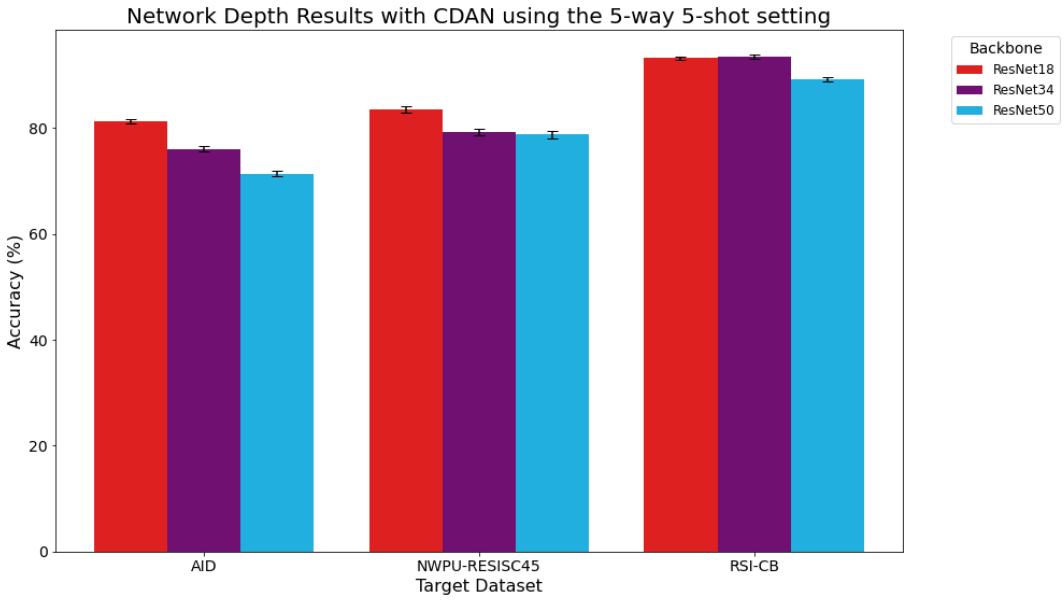


Figure 4.9: CDFS classification network depth test results on different target datasets using CDAN with 5-way 5-shot settings.

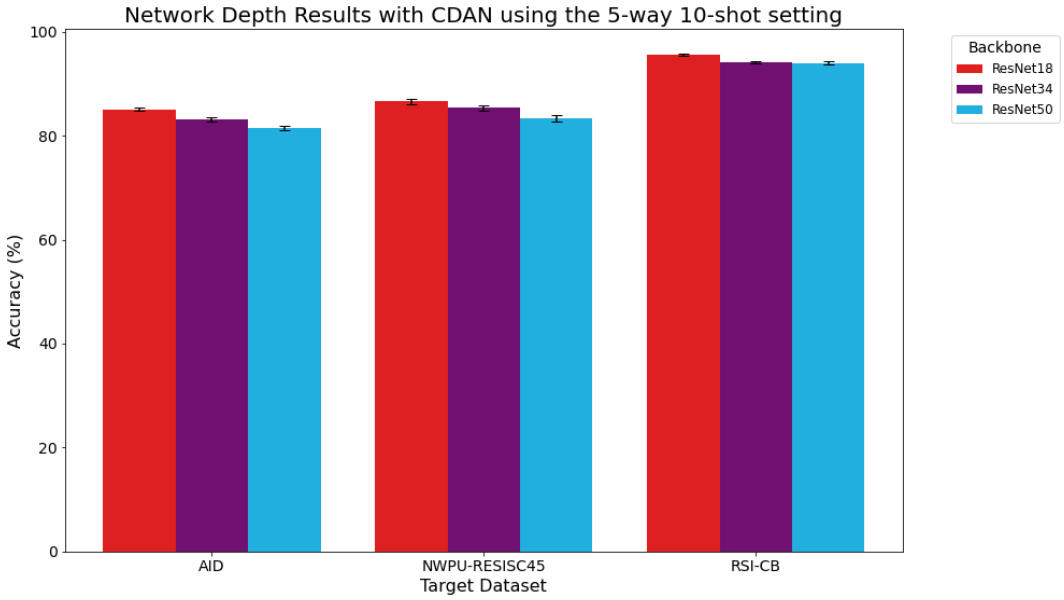


Figure 4.10: CDFS classification network depth test results on different target datasets using CDAN with 5-way 10-shot settings.

4.8 Ablation Study

The results of ablation studies are reviewed and discussed in this section to validate each of the additions for the proposed network that were introduced in chapter 3. This study is important to understand the contribution of each component to the overall method to justify the complexity of the network. It also aids in proving the hypothesis and provides deeper insights into how the proposed network works. The first ablation study examines the impact of the Cross-Domain Region MixUp addition on a network, and then compares it to a network without the advanced augmentation technique. The second ablation study examines the impact of the Dynamically Focused Attention Module addition in a network and compares it to a network without the attention modules. Lastly, the main ablation examines the cumulative effect of the three additions proposed in this dissertation, where each addition is introduced to the network in the order they were discussed. All ablation studies in this section are carried out in a CDFS classification setting with all of the target datasets (AID, NWPU-RESISC45 and RSI-CB).

4.8.1 Effect of Cross-Domain Region MixUp

In this section, the effect of the Cross-Domain Region MixUp (CDRM) addition is discussed and reviewed in isolation to determine the contribution and value of the augmentation technique. Table 4.4 represents the results of the CDFS classification results on ProtoNet with and without the addition of the CDRM method. The choice of comparison to ProtoNet is because it is the chosen base architecture for CDAN as mentioned in section 3.7.1. Results cover all 5-way K-shot settings, and has results from the three target datasets (AID, NWPU-RESISC45 and RSI-CB).

The results from table 4.4 show that adding CDRM clearly improves the performance of the architecture in a CDFS scenario with AID for all K-shot settings. The improvement of classification accuracy with CDRM is significant, at more than 5%, with the 5-shot and 10-shot settings. These results imply that CDRM leads to a network being more effective when there is an increase in sample size (K). Similar trends are observed in table 4.4 on the NWPU-RESISC45 and RSI-CB datasets. These findings indicate the effectiveness and value of the CDRM on its own and its contribution to CDAN in a CDFS setting with remote sensing datasets. The CDRM is a cross-domain augmentation method by nature and this study has demonstrated its ability to address domain adaptation challenges and help the network learn improved domain invariant features, which leads to better generalisation and performance during testing.

Table 4.4: Ablation study test results highlighting the effect of CDRM on ProtoNet's classification accuracy in 5-way K-shot tasks for different target datasets.

Dataset	Architecture	1-Shot	5-Shot	10-Shot
AID	ProtoNet without CDRM	48.52% ± 0.64%	69.68% ± 0.57%	77.24% ± 0.45%
	ProtoNet with CDRM	**48.60% ± 0.77%**	**79.19% ± 0.43%**	**84.19% ± 0.37%**
NWPU-RESISC45	ProtoNet without CDRM	44.40% ± 0.76%	65.11% ± 0.78%	73.96% ± 0.68%
	ProtoNet with CDRM	**58.37% ± 0.91%**	**81.80% ± 0.60%**	**84.80% ± 0.53%**
RSI-CB	ProtoNet without CDRM	67.74% ± 0.69%	90.87% ± 0.36%	92.02% ± 0.33%
	ProtoNet with CDRM	**73.28% ± 0.70%**	**92.09% ± 0.36%**	**94.15% ± 0.28%**

4.8.2 Effect of Dynamically Focused Attention Module

In this section, the effect of the Dynamically Focused Attention Module (DFAM) addition is discussed and reviewed in isolation to determine the contribution and value of the method. Table 4.5 representa the results of the CDFS classification results on ProtoNet with DFAM and without the attention module. The choice of comparison to ProtoNet with DFAM is due to ProtoNet being the chosen base architecture for CDAN, as mentioned in section 3.7.1. Results in the table covers all 5-way K-shot settings, and has results from the three target datasets (AID, NWPU-RESISC45 and RSI-CB).

The experiments mentioned above are quantitative in nature. Therefore, figure 4.11, figure 4.12 and figure 4.13 provide a qualitative view of the different architectures most robust versions (5-way 10-shot) trained in a CDFS classification setting and then tested on unseen novel classes. The results are generated using Grad-CAM [Selvaraju *et al.* 2017]. It is first applied to an image from an unseen class, then the original image is then overlaid with the generated visualisations. Grad-CAM is able to use gradients to determine the important spatial regions within convolutional layers and is therefore helpful with indicating the effect of attention within the CDFS classification setting.

Table 4.5: Ablation study test results highlighting the effect of DFAM on ProtoNet's classification accuracy in 5-way K-shot tasks for different target datasets.

Dataset	Architecture	1-Shot	5-Shot	10-Shot
AID	ProtoNet without DFAM	**48.52% ± 0.64%**	69.68% ± 0.57%	77.64% ± 0.45%
	ProtoNet with DFAM	45.49% ± 0.65%	**72.65% ± 0.50%**	**77.82% ± 0.48%**
NWPU-RESISC45	ProtoNet without DFAM	**44.40% ± 0.76%**	65.11% ± 0.78%	73.96% ± 0.68%
	ProtoNet with DFAM	44.08% ± 0.78%	**67.41% ± 0.75%**	**74.86% ± 0.65%**
RSI-CB	ProtoNet without DFAM	**67.74% ± 0.69%**	**90.87% ± 0.36%**	92.02% ± 0.33%
	ProtoNet with DFAM	65.25% ± 0.73%	88.91% ± 0.42%	**93.75% ± 0.29%**

The results from table 4.5 on AID indicate that the ProtoNet without the attention modules has the best performance with the 1-shot setting. This implies that the attention actually degrades performance within a CDFS scenario for the 1-shot setting. The finding is not surprising as it was observed in the main experiments and results when

comparing the regular ProtoNet to CBAM. ProtoNet with DFAM is able to perform better than the regular ProtoNet in the 5-shot and 10-shot settings. These findings indicate that ProtoNet with DFAM is better than a network without attention modules since it can attain superior overall performance.

Results in table 4.5 with the NWPU-RESISC45 dataset shows a similar trend to AID where the ProtoNet without attention performs better than the ProtoNet with DFAM with the 1-shot setting. This is also observed with RSI-CB where DFAM performs worse with the 1-shot and 5-shot setting but performance improves with the 10-shot setting when compared to the regular Protonet. These findings show that while DFAM struggles for performance with the 1-shot setting when compared to the regular ProtoNet across target datasets, it is able to achieve similar or better results across most of the 5-shot and 10-shot settings for CDFS classification. Therefore, DFAM can still be considered a valuable contribution given that it outperforms the regular ProtoNet in most cases when tested in isolation, and that it was initially designed to complement the other additions in CDAN for CDFS scenarios.

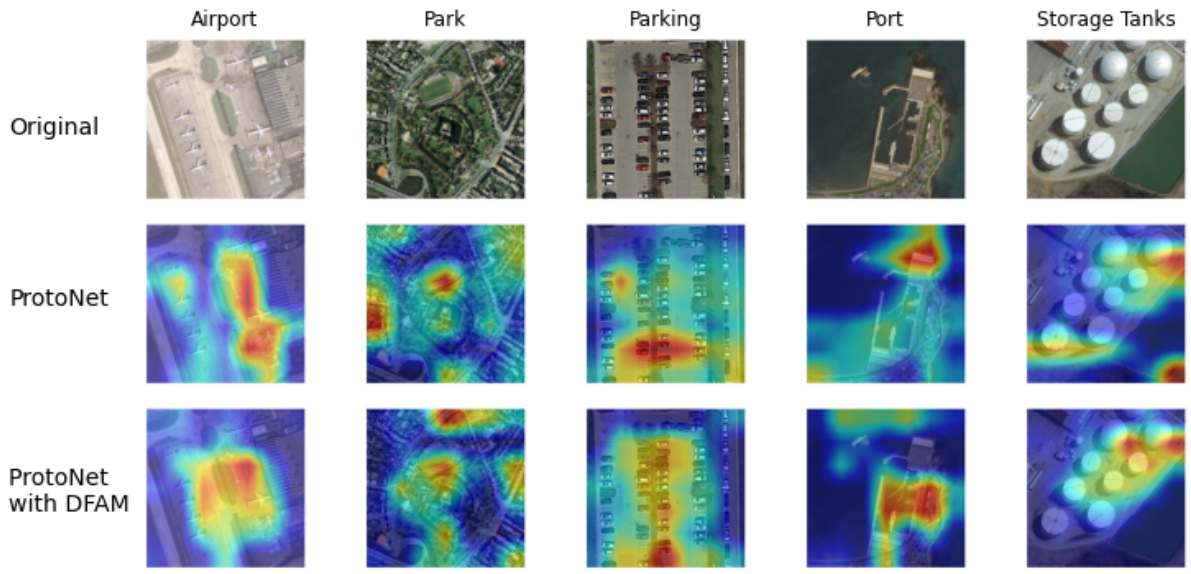


Figure 4.11: Diagram of original scene images and attention maps derived from Grad-Cam using ProtoNet and ProtoNet with DFAM architectures. Classes are chosen from the AID novel/test set.

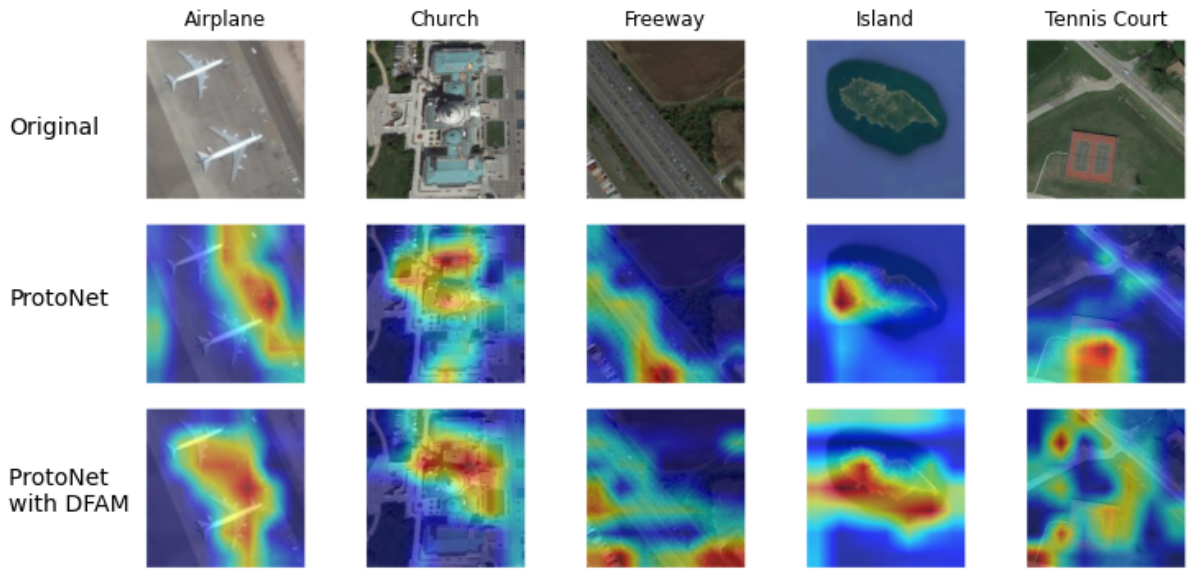


Figure 4.12: Diagram of original scene images and attention maps derived from Grad-Cam using ProtoNet and ProtoNet with DFAM architectures. Classes are chosen from the NWPU-RESISC45 novel/test set.

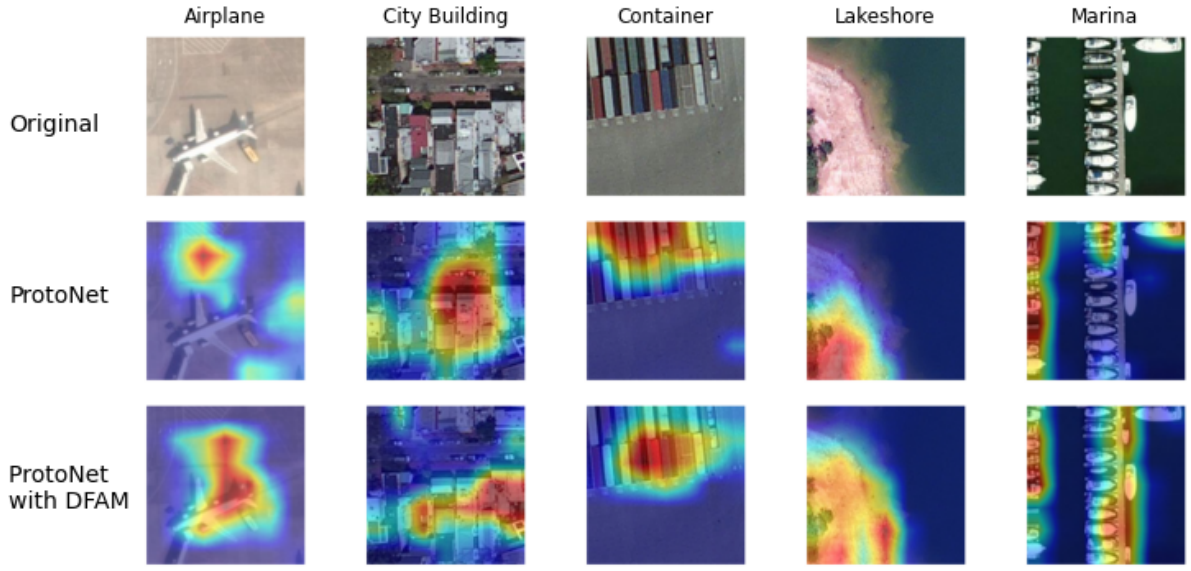


Figure 4.13: Diagram of original scene images and attention maps derived from Grad-Cam using ProtoNet and ProtoNet with DFAM architectures. Classes are chosen from the RSI-CB novel/test set.

The qualitative results in figure 4.11, figure 4.12 and figure 4.13 are generated by passing images from the test classes of the target datasets through the trained architectures and then applying the Grad-CAM method. The method depicts regions where the most focus is placed so the discriminative ability of the architectures can be visualised [Yuan *et al.* 2020]. The intensity of colour indicates the level of importance in the image, where blue implies less attention and red implies more attention within an area based on the presence of relevant objects. All three figures show the original image samples in the first row from the target dataset test classes. The second row consists of the GradCam visualisations overlaid on the original images for the ProtoNet architecture without any attention modules. The third row consists of the GradCam visualisations overlaid on the original images for the ProtoNet architecture with the proposed DFAM.

Figure 4.11, figure 4.12 and figure 4.13 shows that the ProtoNet architecture with DFAM extracts better discriminative information from the scenes in general. The primary objects of interest and additional features are captured to a greater degree compared to results from the ProtoNet without attention modules. It should also be noted that GradCam maps do not have to always tightly match human perception of the target object. A wider coverage in the GradCam heatmap does not automatically mean incorrect attention; it can mean the model is aggregating more features from different parts of the image. Therefore, the network may be attending to other scene features that help distinguish the class. Overall, these results further justify the usage of the proposed attention module and demonstrate that it fulfils its intended purpose, which is to increase the discriminative power of the architecture for the classification of remote sensing images in a CDFS setting.

4.8.3 Effect of all Additions

This section reviews and discusses the results of the primary ablation study. Each of the methods introduced in chapter 3 is added to the base architecture in the order they are introduced to produce the final proposed network, which is then tested across different settings. This will assist in evaluating the impact of each method as they are added to the proposed network. Table 4.6, table 4.7 and table 4.8 represent the results of the CDFS classification results of the main ablation study for all 5-way K-shot settings across each of the target datasets (AID, NWPU-RESISC45 and RSI-CB).

The results from table 4.6 on AID show that the increase in the number of shots leads to an increase in classification accuracy for all architecture options. A clear increase in performance was also observed with the addition of DFAM and the addition of ATAL. This is true for the settings where there is one or five samples available, but once ATAL is added to the architecture with the 10-shot setting, there is an observed decrease in performance from 85.73% to 85.04%. The results from table 4.7 on NWPU-RESISC45 and table 4.8 on RSI-CB have similar trends, but as additions are made to the architecture, there is an apparent increase in terms of performance across all K-shot settings with both datasets.

These ablation study results also justify the usage of the proposed ATAL as it is added to the architecture in table 4.6, table 4.7 and table 4.8. The results demonstrate that ATAL fulfils its intended purpose, which is to improve upon the original cross-entropy loss function and increase classification performance of the architecture by utilising information present in few-shot tasks for the classification of remote sensing images in a CDFS setting.

Overall, these ablation study results indicate that each addition is valuable to the proposed method in a CDFS classification setting with remote sensing imagery. Therefore, each of the additions is essential in proving the hypothesis and justifies the complexity of CDAN.

Table 4.6: CDFS classification accuracy test results for main ablation study with 5-way K-shot settings using AID.

Architecture	**1-Shot**	**5-Shot**	**10-Shot**
ProtoNet+CDRM	48.60% ± 0.77%	79.19% ± 0.43%	84.19% ± 0.37%
ProtoNet+CDRM+DFAM	51.42% ± 0.74%	80.31% ± 0.44%	**85.73% ± 0.35%**
ProtoNet+CDRM+DFAM+ATAL	**55.30% ± 0.76%**	**81.31% ± 0.41%**	85.04% ± 0.34%

Table 4.7: CDFS classification accuracy test results for main ablation study with 5-way K-shot settings using NWPU-RESISC45.

Architecture	**1-Shot**	**5-Shot**	**10-Shot**
ProtoNet+CDRM	58.37% ± 0.91%	81.80% ± 0.60%	84.80% ± 0.53%
ProtoNet+CDRM+DFAM	61.59% ± 0.90%	82.57% ± 0.60%	85.22% ± 0.52%
ProtoNet+CDRM+DFAM+ATAL	**62.06% ± 0.91%**	**83.56% ± 0.59%**	**86.63% ± 0.51%**

Table 4.8: CDFS classification accuracy test results for main ablation study with 5-way K-shot settings using RSI-CB.

Architecture	**1-Shot**	**5-Shot**	**10-Shot**
ProtoNet+CDRM	73.28% ± 0.70%	92.09% ± 0.36%	94.15% ± 0.28%
ProtoNet+CDRM+DFAM	74.37% ± 0.68%	92.74% ± 0.38%	94.77% ± 0.28%
ProtoNet+CDRM+DFAM+ATAL	**76.03% ± 0.71%**	**93.18% ± 0.34%**	**95.51% ± 0.23%**

4.9 Summary

In this chapter, all of the experiments and results are discussed in detail. The main experiments with AID, NWPU-RESISC45 and RSI-CB as the target datasets show that the proposed network (CDAN) is effective in a CDFS classification setting, using prediction accuracy and F1 score metrics for evaluation. CDAN outperforms the traditional

few-shot learning architecture as well as networks with related attention and augmentation methods. Furthermore, it proves that it is able to perform well when a significant domain shift exists and a limited amount of labelled data is available.

There were a number of additional experiments that were conducted. Firstly, the number of classes is varied during testing to show the effect on prediction accuracy on CDAN. Secondly, the experiment for the scaling factor justifies the choice used for CDAN in the main and additional experiments. Thirdly, it is shown that an increase in network depth actually decreases the performance of CDAN in a CDFS classification setting. Lastly, the in-depth ablation study validates the contributions and justifies the complexity of the overall proposed network in this dissertation with quantitative and qualitative results.

Chapter 5

Conclusion

5.1 Research Summary

This Cross-Domain Attention Network (CDAN) is proposed in this dissertation. It is a modified convolutional neural network, created for the purpose of solving challenges associated with CDFS scene classification on remote sensing imagery. These challenges include a significant domain shift, a limited amount of labelled data and the complex characteristics of remote sensing imagery. The proposed network introduces three innovative additions to tackle the challenges by improving overall adaptability and classification performance. Each of the additions will be summarised below.

The first addition is Cross-Domain Region MixUp (CDRM), introduced in section 3.6.2. This is a data augmentation technique by nature, and it focuses on improving adaptability by reducing domain shift and enhancing the ability of the network to generalise across domains in scenarios where there is a limited amount of labelled data available. This is achieved by cutting and pasting relevant regions from the target domain onto the input images from the source domain during training, and it also prevents the occlusion of objects by blending the pasted region with the input image. The effectiveness of the technique is demonstrated with the complete network across three different remote sensing target datasets in the main experiments. Furthermore, the addition of CDRM is justified, and the value of the technique demonstrated with comprehensive ablation studies in section 4.8.1 and section 4.8.3.

The second addition is Dynamically Focused Attention Module (DFAM), introduced in section 3.6.3. The attention module developed in this dissertation focuses on increasing the discriminative power of the network to handle the complexity and variability within the data. This is achieved by improving the ability of the network to focus on important features, through effective use of the channel and spatial information available during the training process. The overall effectiveness of the attention module is demonstrated with the complete network across three different remote sensing target datasets in the

main experiments. Additionally, DFAM is evaluated and justified with comprehensive ablation studies in section 4.8.2 and section 4.8.3.

The third addition is Adaptive Task Aware Loss (ATAL), introduced in section 3.6.4. The loss function developed in this dissertation focuses on improving the adaptability of the network to new categories with minimal labelled remote sensing data. This is achieved by extending the existing cross-entropy loss to account for individual task complexity and improving the focus on intra-class characteristics of the data during training, which enables more balanced learning and network robustness in a CDFS scenario. The overall effectiveness of the loss function is demonstrated with the complete network across three different remote sensing target datasets in the main experiments. Additionally, ATAL is evaluated and justified with a comprehensive ablation study in section 4.8.3.

The dissertation showcases a thorough evaluation of CDAN, which demonstrates that it outperforms traditional few-shot methods as well as related attention and augmentation methods in CDFS classification tasks that involve remote sensing imagery. The main experiments show that supervised learning approaches do not perform well in limited labelled data settings, while regular attention and augmentation methods are not suitable in a CDFS setting due to potential overfitting on the source domain. Additional experiments in section 4.5 demonstrate the number of classes in a task determines the difficulty of the task, where more classes lead to a decrease in accuracy and indicate more challenging tasks. In section 4.7, the increase in network depth was also shown to decrease classification performance in a CDFS setting, with shallow networks proving to be more effective.

The key finding from this dissertation demonstrates that CDAN is highly effective in scenarios where there is a significant domain shift between the source domain dataset (natural imagery) and the target domain dataset (remote sensing imagery). Additionally, the network is able to perform well in an environment where access to labelled data is limited, and this characteristic makes it valuable in the remote sensing field where large labelled datasets are not readily available. Therefore, this dissertation and the solution it has introduced has validated the hypothesis stated in section 3.2.

In conclusion, this dissertation has introduced a robust solution to the problem of CDFS scene classification for remote sensing imagery, which has been validated through extensive experimentation. The results confirm the effectiveness of modifying a CNN with an advanced domain-level data augmentation technique, attention modules and an improved loss function in a CDFS learning paradigm, which provides practical applications in a field with limited literature.

5.2 Limitations of the Study

While this dissertation has empirically validated the research hypothesis, and it can be concluded that incorporating CDAN into the CDFS learning paradigm is valuable, the limitations of the research need to be restated given potential applications in domains not explored in this dissertation. Firstly, this dissertation does not seek to evaluate the efficacy of CDRM, DFAM, or ATAL beyond the realm of CNNs. While it is possible to make reasonable assumptions about how these techniques may generalise in other contexts, such considerations are beyond the scope of this study.

Secondly, the focus of the dissertation was placed on CDFS classification tasks that involved a significant domain shift between the source domain dataset (natural imagery) and target domain dataset (remote sensing). The scope of the existing focus is already substantial and requires considerable time and effort to produce meaningful results. Additionally, it is a challenging task to set up and run these types of experiments due to the limited access to resources. As a result, this dissertation does not consider other cases or datasets from unseen domains.

Lastly, this dissertation evaluates the proposed network with three different target remote sensing datasets. The influence of the characteristics of the target datasets, with varying resolutions and properties, affects the performance of the network in CDFS classification and can significantly affect the results achieved during testing. These aspects pose challenges that impact the network's performance and have not been fully explored in this work.

5.3 Future Work

The dissertation provides a foundation for future research to consider addressing some of the limitations discussed above, which may include extending methods and ideas to non-CNN architecture, data from different domains and their impact on CDFS learning.

Appendix A

Additional Information

A.1 Source Dataset Split

Table A.1: Dataset split for Mini-ImageNet.

Data Subset	Classes
Base	n01532829, n01558993, n01704323, n01749939, n01770081, n01843383, n01855672, n01910747, n01930112, n01981276, n02074367, n02089867, n02091244, n02091831, n02099601, n02101006, n02105505, n02108089, n02108551, n02108915, n02110063, n02110341, n02111277, n02113712, n02114548, n02116738, n02120079, n02129165, n02138441, n02165456, n02174001, n02219486, n02443484, n02457408, n02606052, n02687172, n02747177, n02795169, n02823428, n02871525, n02950826, n02966193, n02971356, n02981792, n03017168, n03047690, n03062245, n03075370, n03127925, n03146219, n03207743, n03220513, n03272010, n03337140, n03347037, n03400231, n03417042, n03476684, n03527444, n03535780, n03544143, n03584254, n03676483, n03770439, n03773504, n03775546, n03838899, n03854065, n03888605, n03908618, n03924679, n03980874, n03998194, n04067472, n04146614, n04149813, n04243546, n04251144, n04258138, n04275548
Val	n04522168, n04604644, n04509417, n04296562, n09246464, n07613480, n04389033, n06794110, n04443257, n04612504
Novel	n13054560, n04515003, n04435653, n04596742, n07584110, n09256479, n07747607, n13133613, n04418357, n07697537

A.2 Target Dataset Splits

Table A.2: Dataset split for AID.

Data Subset	Classes
Base	BaseballField, Beach, Church, Commercial, DenseResidential, Desert, Forest, Industrial, Meadow, MediumResidential, Mountain, Playground, Pond, RailwayStation, Resort, River, School, Stadium
Val	BareLand, Center, Bridge, SparseResidential, Square, Viaduct
Novel	StorageTanks, Park, Farmland, Parking, Airport, Port

Table A.3: Dataset split for NWPU-RESISC45.

Data Subset	Classes
Base	airport, baseball diamond, beach, bridge, chaparral, desert, forest, golf course, industrial area, intersection, lake, meadow, medium residential, mountain, palace, parking lot, railway, rectangular farmland, river, roundabout, sea ice, ship, snowberg, sparse residential, stadium, terrace, thermal power station
Val	basketball court, cloud, dense residential, ground track field, railway station, runway, storage tank, wetland, overpass
Novel	commercial area, freeway, circular farmland, tennis court, harbor, church, island, airplane, mobile home park

Table A.4: Dataset split for RSI-CB.

Data Subset	Classes
Base	avenue, bare land, bridge, coastline, dam, desert, forest, green farmland, highway, mangrove, mountain, parkinglot, pipeline, residents, river, sapling, sea, shrubwood, snow mountain, storage room, town
Val	artificial grassland, crossroads, dry farm, river protection forest, sparse forest, sandbeach, hirst
Novel	container, stream, city building, marina, airplane, airport runway, lakeshore

A.3 ResNet18 Structure

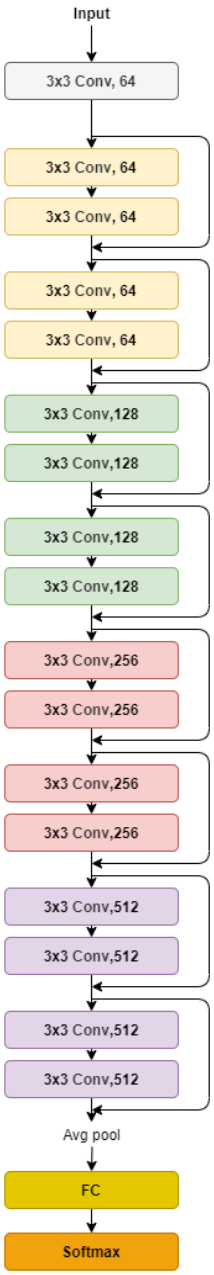


Figure A.1: The ResNet18 CNN structure.

A.4 Architecture Parameters

Table A.5: Total number of learnable parameters for different architectures.

Architecture	**Learnable Parameters**
Baseline	11,176,512
MatchingNet	11,176,512
RelationNet	11,176,512
ProtoNet	11,176,512
CBAM	11,266,392
MixUp	11,176,512
CutMix	11,176,512
Cross-Domain CutMix	11,176,512
CDAN (ours)	11,355,560

References

[Al-Doski *et al.* 2013] Jwan Al-Doski, Shattri B Mansorl, and Helmi Zulhaidi Mohd Shafri. Image classification in remote sensing. *Department of Civil Engineering, Faculty of Engineering, University Putra, Malaysia*, 3(10), 2013.

[Alajaji *et al.* 2020] Dalal Alajaji, Haikel S Alhichri, Nassim Ammour, and Naif Alajlan. Few-shot learning for remote sensing scene classification. In *2020 Mediterranean and Middle-East Geoscience and Remote Sensing Symposium (M2GARSS)*, pages 81–84. IEEE, 2020.

[Bengio *et al.* 2017] Yoshua Bengio, Ian Goodfellow, and Aaron Courville. *Deep learning*, volume 1. MIT press Cambridge, MA, USA, 2017.

[Berrada *et al.* 2018] Leonard Berrada, Andrew Zisserman, and M Pawan Kumar. Smooth loss functions for deep top-k classification. *arXiv preprint arXiv:1802.07595*, 2018.

[Chaib *et al.* 2017] Souleyman Chaib, Huan Liu, Yanfeng Gu, and Hongxun Yao. Deep feature fusion for vhr remote sensing scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(8):4775–4784, 2017.

[Chen *et al.* 2018] Jingbo Chen, Chengyi Wang, Zhong Ma, Jiansheng Chen, Dongxu He, and Stephen Ackland. Remote sensing scene classification based on convolutional neural networks pre-trained using attention-guided sparse filters. *Remote Sensing*, 10(2):290, 2018.

[Chen *et al.* 2019] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. *arXiv preprint arXiv:1904.04232*, 2019.

[Cheng *et al.* 2017] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

[Cheng *et al.* 2018] Gong Cheng, Ceyuan Yang, Xiwen Yao, Lei Guo, and Junwei Han. When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative cnns. *IEEE transactions on geoscience and remote sensing*, 56(5):2811–2821, 2018.

[Cheng *et al.* 2021] Yuan-Chia Cheng, Ci-Siang Lin, Fu-En Yang, and Yu-Chiang Frank Wang. Few-shot classification in unseen domains by episodic meta-learning across visual domains. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 434–438. IEEE, 2021.

[Demirkaya *et al.* 2020] Ahmet Demirkaya, Jiasi Chen, and Samet Oymak. Exploring the role of loss functions in multiclass classification. In *2020 54th annual conference on information sciences and systems (ciss)*, pages 1–5. IEEE, 2020.

[Deng *et al.* 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE, 2009.

[DeVries and Taylor 2017] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[Ehteshami Bejnordi and Krestel 2020] Ali Ehteshami Bejnordi and Ralf Krestel. Dynamic channel and layer gating in convolutional neural networks. In *KI 2020: Advances in Artificial Intelligence: 43rd German Conference on AI, Bamberg, Germany, September 21–25, 2020, Proceedings 43*, pages 33–45. Springer, 2020.

[Filipi Gonçalves dos Santos and Papa 2022] Claudio Filipi Gonçalves dos Santos and João Paulo Papa. Avoiding overfitting: A survey on regularization methods for convolutional neural networks. *arXiv e-prints*, pages arXiv–2201, 2022.

[Finn *et al.* 2017] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017.

[Gao *et al.* 2018] Xitong Gao, Yiren Zhao, Łukasz Dudziak, Robert Mullins, and Cheng-zhong Xu. Dynamic channel pruning: Feature boosting and suppression. *arXiv preprint arXiv:1810.05331*, 2018.

[Gong *et al.* 2018] Zhiqiang Gong, Ping Zhong, Yang Yu, and Weidong Hu. Diversity-promoting deep structural metric learning for remote sensing scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 56(1):371–390, 2018.

[Gu *et al.* 2018] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, et al. Recent advances in convolutional neural networks. *Pattern Recognition*, 77:354–377, 2018.

[Guo *et al.* 2019] Yiyou Guo, Jinsheng Ji, Xiankai Lu, Hong Huo, Tao Fang, and Deren Li. Global-local attention network for aerial scene classification. *IEEE Access*, 7:67200–67212, 2019.

[Guo *et al.* 2020] Yunhui Guo, Noel C Codella, Leonid Karlinsky, James V Codella, John R Smith, Kate Saenko, Tajana Rosing, and Rogerio Feris. A broader study of cross-domain few-shot learning. In *European Conference on Computer Vision*, pages 124–141. Springer, 2020.

[He *et al.* 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE, 2016.

[Hemati *et al.* 2021] MohammadAli Hemati, Mahdi Hasanlou, Masoud Mahdianpari, and Fariba Mohammadimanesh. A systematic review of landsat data for change detection applications: 50 years of monitoring the earth. *Remote sensing*, 13(15):2869, 2021.

[Hossain and Sajib 2019] Md Anwar Hossain and Md Shahriar Alam Sajib. Classification of image using convolutional neural network (cnn). *Global Journal of Computer Science and Technology*, 19(2), 2019.

[Hua *et al.* 2019] Weizhe Hua, Yuan Zhou, Christopher M De Sa, Zhiru Zhang, and G Edward Suh. Channel gating neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[Huang *et al.* 2019] Xin Huang, Xiaopeng Han, Song Ma, Tianjia Lin, and Jianya Gong. Monitoring ecosystem service change in the city of shenzhen by the use of high-resolution remotely sensed imagery and deep learning. *Land Degradation & Development*, 30(12):1490–1501, 2019.

[Ji *et al.* 2022] Hong Ji, Zhi Gao, Yongjun Zhang, Yu Wan, Can Li, and Tiancan Mei. Few-shot scene classification of optical remote sensing images leveraging calibrated pretext tasks. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022.

[Kim *et al.* 2021] Yechan Kim, Younkwan Lee, and Moongu Jeon. Imbalanced image classification with complement cross entropy. *Pattern Recognition Letters*, 151:33–40, 2021.

[King *et al.* 2016] Josh King, Vayu Kishore, and Filippo Ranalli. Scene classification with convolutional neural networks. *cs231n. stanford. edu*, 2016.

[Kong *et al.* 2021] Jiayuan Kong, Yurong Gao, Yanjun Zhang, Huimin Lei, Yao Wang, and Hesheng Zhang. Improved attention mechanism and residual network for remote sensing image scene classification. *Ieee Access*, 9:134800–134808, 2021.

[Krizhevsky *et al.* 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, volume 25, pages 1097–1105, 2012.

[Li and Wang 2020] Xiaobin Li and Weiqiang Wang. Learning discriminative features via weights-biased softmax loss. *Pattern Recognition*, 107:107405, 2020.

[Li *et al.* 2017a] Erzhu Li, Junshi Xia, Peijun Du, Cong Lin, and Alim Samat. Integrating multilayer features of convolutional neural networks for remote sensing scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(10):5653–5665, 2017.

[Li *et al.* 2017b] Haifeng Li, Xin Dou, Chao Tao, Zhixiang Hou, Jie Chen, Jian Peng, Min Deng, and Ling Zhao. Rsi-cb: A large scale remote sensing image classification benchmark via crowdsource data. *arXiv preprint arXiv:1705.10450*, 2017.

[Li *et al.* 2020a] Haifeng Li, Zhenqi Cui, Zhiqing Zhu, Li Chen, Jiawei Zhu, Haozhe Huang, and Chao Tao. Rs-metanet: Deep meta metric learning for few-shot remote sensing scene classification. *arXiv preprint arXiv:2009.13364*, 2020.

[Li *et al.* 2020b] Lingjun Li, Junwei Han, Xiwen Yao, Gong Cheng, and Lei Guo. Dlamatchnet for few-shot remote sensing image scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[Li *et al.* 2020c] Wenmei Li, Ziteng Wang, Yu Wang, Jiaqi Wu, Juan Wang, Yan Jia, and Guan Gui. Classification of high-spatial-resolution remote sensing scenes method using transfer learning and deep convolutional neural network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:1986–1995, 2020.

[Li *et al.* 2020d] Xirong Li, Fangling Pu, Rui Yang, Rong Gui, and Xin Xu. Amn: Attention metric network for one-shot remote sensing image scene classification. *Remote Sensing*, 12(24):4046, 2020.

[Li *et al.* 2021] Zhaokui Li, Ming Liu, Yushi Chen, Yimin Xu, Wei Li, and Qian Du. Deep cross-domain few-shot learning for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[Li *et al.* 2023] Can Li, He Chen, Yin Zhuang, and Shanghang Zhang. Uncertainty-aware dynamic learning for cross-domain few-shot scene classification from remote sensing imagery. In *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*, pages 5778–5781. IEEE, 2023.

[Lin *et al.* 2017] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988. IEEE, 2017.

[Liu *et al.* 2018] Yanfei Liu, Yanfei Zhong, and Qianqing Qin. Scene classification based on multiscale convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 56(12):7109–7121, 2018.

[Liu *et al.* 2020] Jie Liu, Wenjie Zhang, Yuting Tang, Jie Tang, and Gangshan Wu. Residual feature aggregation network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2359–2368. IEEE, 2020.

[Lu *et al.* 2017] Xiaoqiang Lu, Xiangtao Zheng, and Yuan Yuan. Remote sensing scene classification by unsupervised representation learning. *IEEE Transactions on Geoscience and Remote Sensing*, 55(9):5148–5157, 2017.

[Ma *et al.* 2015] Yan Ma, Haiping Wu, Lizhe Wang, Bormin Huang, Rajiv Ranjan, Albert Zomaya, and Wei Jie. Remote sensing big data computing: Challenges and opportunities. *Future Generation Computer Systems*, 51:47–60, 2015.

[Ma *et al.* 2019a] Chenhui Ma, Xiaodong Mu, and Dexuan Sha. Multi-layer feature fusion of convolutional neural networks for remote sensing scene classification. *IEEE Access*, 7:121685–121694, 2019.

[Ma *et al.* 2019b] Dongao Ma, Ping Tang, and Lijun Zhao. Siftinggan: Generating and sifting labeled samples to improve the remote sensing image scene classification baseline in vitro. *IEEE Geoscience and Remote Sensing Letters*, 16(7):1046–1050, 2019.

[Ma *et al.* 2022] Ailong Ma, Ning Yu, Zhuo Zheng, Yanfei Zhong, and Liangpei Zhang. A supervised progressive growing generative adversarial network for remote sensing image scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–18, 2022.

[Mekhalfi *et al.* 2015] Mohamed L Mekhalfi, Farid Melgani, Yakoub Bazi, and Naif Alajlan. Land-use classification with compressive sensing multifeature fusion. *IEEE Geoscience and Remote Sensing Letters*, 12(10):2155–2159, 2015.

[Nakamura *et al.* 2022] Yuzuru Nakamura, Yasunori Ishii, Yuki Maruyama, and Takayoshi Yamashita. Few-shot adaptive object detection with cross-domain cutmix. In *Proceedings of the Asian Conference on Computer Vision*, pages 1350–1367. Springer, 2022.

[Powers 2020] David MW Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*, 2020.

[Ren *et al.* 2018] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B. Tenenbaum, Hugo Larochelle, and Richard S. Zemel. Meta-learning for semi-supervised few-shot classification. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, 2018.

[Rostami *et al.* 2019] Mohammad Rostami, Soheil Kolouri, Eric Eaton, and Kyungnam Kim. Sar image classification using few-shot cross-domain transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2019.

[Rubner *et al.* 2000] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000.

[Sa *et al.* 2022] Liangbing Sa, Chongchong Yu, Xianqin Ma, Xia Zhao, and Tao Xie. Attentive fine-grained recognition for cross-domain few-shot classification. *Neural Computing and Applications*, 34(6):4733–4746, 2022.

[Selvaraju *et al.* 2017] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626. IEEE, 2017.

[Shang *et al.* 2018] Ronghua Shang, Jiaming Wang, Licheng Jiao, Rustam Stolkin, Biao Hou, and Yangyang Li. Sar targets classification based on deep memory convolution neural networks and transfer parameters. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(8):2834–2846, 2018.

[Singh and Bruzzone 2021] Abhishek Singh and Lorenzo Bruzzone. Sigan: Spectral index generative adversarial network for data augmentation in multispectral remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021.

[Snell *et al.* 2017] Jake Snell, Kevin Swersky, and Richard S Zemel. Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175*, 2017.

[Sun *et al.* 2021] Xian Sun, Bing Wang, Zhirui Wang, Hao Li, Hengchao Li, and Kun Fu. Research progress on few-shot learning for remote sensing image interpretation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:2387–2402, 2021.

[Sung *et al.* 2018] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208. IEEE, 2018.

[Tane *et al.* 2018] Zachary Tane, Dar Roberts, Alexander Koltunov, Stuart Sweeney, and Carlos Ramirez. A framework for detecting conifer mortality across an ecoregion using high spatial resolution spaceborne imaging spectroscopy. *Remote Sensing of Environment*, 209:195–210, 2018.

[Tong *et al.* 2020] Wei Tong, Weitao Chen, Wei Han, Xianju Li, and Lizhe Wang. Channel-attention-based densenet network for remote sensing image scene classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:4121–4132, 2020.

[Torrey and Shavlik 2010] Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global, 2010.

[Tseng *et al.* 2020] Hung-Yu Tseng, Hsin-Ying Lee, Jia-Bin Huang, and Ming-Hsuan Yang. Cross-domain few-shot classification via learned feature-wise transformation. *arXiv preprint arXiv:2001.08735*, 2020.

[Uddin *et al.* 2020] AFM Uddin, Mst Monira, Wheemyung Shin, TaeChoong Chung, Sung-Ho Bae, et al. Saliencymix: A saliency guided data augmentation strategy for better regularization. *arXiv preprint arXiv:2006.01791*, 2020.

[Vinyals *et al.* 2016] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. *arXiv preprint arXiv:1606.04080*, 2016.

[Wang and others 2016] Xiaogang Wang et al. Deep learning in object recognition, detection, and segmentation. *Foundations and Trends® in Signal Processing*, 8(4):217–382, 2016.

[Wang *et al.* 2017] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164. IEEE, 2017.

[Wang *et al.* 2018a] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5265–5274. IEEE, 2018.

[Wang *et al.* 2018b] Yuntao Wang, Albert S Chen, Guangtao Fu, Slobodan Djordjević, Chi Zhang, and Dragan A Savić. An integrated framework for high-resolution urban flood modelling considering multiple information sources and urban features. *Environmental modelling & software*, 107:85–95, 2018.

[Wang *et al.* 2018c] Zhaocheng Wang, Lan Du, Jiashun Mao, Bin Liu, and Dongwen Yang. Sar target detection based on ssd with data augmentation and transfer learning. *IEEE Geoscience and Remote Sensing Letters*, 16(1):150–154, 2018.

[Wang *et al.* 2019] Hao Wang, Chao Tao, Ji Qi, Haifeng Li, and YuQi Tang. Semi-supervised variational generative adversarial networks for hyperspectral image classification. In *IGARSS 2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 9792–9794. IEEE, 2019.

[Wang *et al.* 2020] Qi Wang, Yue Ma, Kun Zhao, and Yingjie Tian. A comprehensive survey of loss functions in machine learning. *Annals of Data Science*, pages 1–26, 2020.

[Wang *et al.* 2023a] Wenjian Wang, Lijuan Duan, Yuxi Wang, Junsong Fan, Zhi Gong, and Zhaoxiang Zhang. A survey of deep visual cross-domain few-shot learning. *arXiv preprint arXiv:2303.09253*, 2023.

[Wang *et al.* 2023b] Wenzhen Wang, Fang Liu, Jia Liu, and Liang Xiao. Cross-domain few-shot hyperspectral image classification with class-wise attention. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–18, 2023.

[Weinberger and Saul 2009] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of machine learning research*, 10(2), 2009.

[Woo *et al.* 2018] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19. Springer, 2018.

[Xia *et al.* 2017] Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, 2017.

[Xu *et al.* 2021a] Hanqing Xu, Ming Yang, Liuyuan Deng, Yeqiang Qian, and Chunxiang Wang. Neutral cross-entropy loss based unsupervised domain adaptation for semantic segmentation. *IEEE Transactions on Image Processing*, 30:4516–4525, 2021.

[Xu *et al.* 2021b] Jingyi Xu, Hieu Le, Mingzhen Huang, ShahRukh Athar, and Dimitris Samaras. Variational feature disentangling for fine-grained few-shot classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 8812–8821. IEEE, 2021.

[Yan *et al.* 2020] Peiyao Yan, Feng He, Yajie Yang, and Fei Hu. Semi-supervised representation learning for remote sensing image classification based on generative adversarial networks. *IEEE Access*, 8:54135–54144, 2020.

[Yu *et al.* 2017] Xingrui Yu, Xiaomin Wu, Chunbo Luo, and Peng Ren. Deep learning in remote sensing scene classification: a data augmentation enhanced convolutional neural network framework. *GIScience & Remote Sensing*, 54(5):741–758, 2017.

[Yuan *et al.* 2020] Zhengwu Yuan, Wendong Huang, Lin Li, and Xiaobo Luo. Few-shot scene classification with multi-attention deepemd network in remote sensing. *IEEE Access*, 9:19891–19901, 2020.

[Yun *et al.* 2019] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 6023–6032. IEEE, 2019.

[Zagoruyko and Komodakis 2016] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

[Zeiler and Fergus 2014] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer, 2014.

[Zeng *et al.* 2021] Delu Zeng, Minyu Liao, Mohammad Tavakolian, Yulan Guo, Bolei Zhou, Dewen Hu, Matti Pietikäinen, and Li Liu. Deep learning for scene classification: A survey. *arXiv preprint arXiv:2101.10531*, 2021.

[Zhang *et al.* 2017] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[Zhang *et al.* 2020a] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12203–12213. IEEE, 2020.

[Zhang *et al.* 2020b] Lamei Zhang, Siyu Zhang, Bin Zou, and Hongwei Dong. Unsupervised deep representation learning and few-shot classification of polsar images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2020.

[Zhang *et al.* 2021] Pei Zhang, Yunpeng Bai, Dong Wang, Bendu Bai, and Ying Li. Few-shot classification of aerial scene images via meta-learning. *Remote Sensing*, 13(1):108, 2021.

[Zhong *et al.* 2020] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13001–13008. AAAI Press, 2020.

Declaration
University of the Witwatersrand, Johannesburg
School of Computer Science and Applied Mathematics
SENATE PLAGIARISM POLICY

I, Christopher Wayne Pillay, (Student number: 1362077) am a student registered for Master of Science (Dissertation) in the year 2025.

I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that ALL the work submitted for assessment for the above course is my own unaided work except where I have explicitly indicated otherwise.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this in not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.

Signature:  **Date**: 11/04/2025