BEYOND 2D AND MORE: INTERPRETING REMOTE SENSING IMAGE CLASSIFICATION METHODS VIA EXPLAINABLE ARTIFICIAL INTELLIGENCE

by
DEREN EGE TURAN

Submitted to the Graduate School of Engineering and Natural Sciences
in partial fulfilment of
the requirements for the degree of Master of Science

Sabancı University
December 2023

ABSTRACT

BEYOND 2D AND MORE: INTERPRETING REMOTE SENSING IMAGE CLASSIFICATION METHODS VIA EXPLAINABLE ARTIFICIAL INTELLIGENCE

DEREN EGE TURAN

COMPUTER SCIENCE AND ENGINEERING MSc THESIS, DECEMBER 2023

Thesis Supervisor: Assoc. Prof. Erchan Aptoula

Keywords: Explainable artificial intelligence, interpretability, hyperspectral images, GradCam, GradCam++, Guided Backpropagation, domain generalization

Within the hyperspectral remote sensing image classification research area, this thesis delves into the challenges of explaining the decision-making process of deep-learning models. The focus is on the integration of three prominent explainable artificial intelligence methods, namely Grad-CAM, Grad-CAM++, and Guided Backpropagation. These methods have been employed in order to comprehend the decision-making process of a typical convolutional neural network model during spatial-spectral hyperspectral image classification. The conducted experiments investigate the impact of varying pixel patch sizes on spatial attention and the significance of individual spectral bands in the classification process. This thesis sheds light on the behavior of convolutional neural networks in the spatial-spectral context, providing a deeper understanding of how these models respond to changes in hyperspectral data. Furthermore, the study analyzes the relative advantages and limitations of the employed explainability techniques —Grad-CAM, Grad-CAM++, and Guided Backpropagation— in explaining the decision-making processes of the convolutional neural network model. In conclusion, the results provide both deeper interpretations of the behavior of convolutional neural networks as well as a comparative performance analysis of explainability techniques.

ÖZET

2 BOYUTTAN DAHA FAZLASI: AÇIKLANABİLİR YAPAY ZEKA ARACILIĞIYLA UZAKTAN ALGILAMA GÖRÜNTÜ SINIFLANDIRMA YÖNTEMLERİNİN YORUMLANMASI

DEREN EGE TURAN

BİLGİSAYAR BİLİMİ VE MÜHENDİSLİĞİ YÜKSEK LİSANS TEZİ, ARALIK 2023

Tez Danışmanı: Doç. Dr. Erchan Aptoula

Anahtar Kelimeler: Açıklanabilir yapay zeka, yorumlanabilirlik, hiperspektral görüntüler, GradCam, GradCam++, Yönlendirilmiş Geriye Yayılım, alan genelleme

Bu çalışma hiperspektral uzaktan algılama görüntü sınıflandırma araştırma alanı içinde, derin öğrenme modellerinin karar verme sürecini açıklamanın zorluklarına odaklanmaktadır. Odak noktası, üç önemli açıklanabilir yapay zeka yönteminin uygulanması üzerinedir; bunlar GradCAM, GradCAM++ ve Yönlendirilmiş Geriye Yayılım'dır. Bu yöntemler, standart bir evrişimli sinir ağı modelinin uzamsal-spektral hiperspektral görüntü sınıflandırma sürecindeki karar verme sürecini anlamak için kullanılmıştır. Gerçekleştirilen deneyler, piksel yama boyutlarının uzamsal dikkat üzerindeki etkisini ve sınıflandırma sürecinde spektral bantların önemini incelemektedir. Bu çalışma, evrişimli sinir ağlarının uzamsal-spektral bağlamdaki davranışını aydınlatarak, bu modellerin hiperspektral verilerdeki değişikliklere nasıl yanıt verdiğine dair daha derin bir anlayış sağlamaktadır. Ek olarak, bu çalışma kullanılan açıklanabilirlik tekniklerinin -GradCAM, GradCAM++ ve Yönlendirilmiş Geriye Yayılım- karar verme süreçlerini açıklama konusundaki göreceli avantajlarını ve sınırlamalarını analiz etmektedir. Özet olarak, elde edilen sonuçlar, hem evrişimli sinir ağlarının davranışıyla ilgili daha derin yorumlar sunmakta hem de karşılaştırmalı olarak açıklanabilirlik tekniklerinin performansını aktarmaktadır.

ACKNOWLEDGEMENTS

I was on the brink of giving up on my master's degree right before the fall semester of 2022-2023. After taking a single class on computer vision, I changed my mind. With that, I am deeply grateful to my thesis advisor, Erchan Aptoula, for his endless enthusiasm for learning and teaching, as well as his continuous support and guidance throughout my time as his student. He not only helped me rediscover my passion for research and programming, which I thought was gone for good, but also showed me that the best research environment happens in a mentally peaceful place. If I ever consider giving up again, I will recall these 1.5 years.

I also want to thank my friends in VPA Lab, including Berker who supported me until the very last day, even though he was at a Christmas gathering, Ali, Mekan, Seçilay, and Semih for providing a cozy, homelike lab experience.

Finally, special thanks to our family's little ones, Ruby, after whom I named my first research computer and deeply miss, and Luca. Their presence has made me feel the love and joy passed on without words, and for that, I am deeply grateful.

For my family...

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATONS

1D One-dimensional . . . 7

2D Two-dimensional . . . 19, 20, 28, 29

3D Three-dimensional . . . 5, 15, 19, 20, 28, 29

AI Artificial Intelligence . . . 3, 4

CAM Class Activation Map . . . 5, 7, 10, 17, 22

CNN Convolutional Neural Networks . . . 5, 10, 14, 15, 17, 18, 19, 20, 28, 52, 54

CondCAD Conditional Contrastive Adversarial Domain . . . 56

Grad-CAM++ Generalized Gradient-weighted Class Activation Mapping iv, xi, xii, 1, 2, 3, 4, 6, 7, 12, 13, 14, 21, 22, 25, 29, 30, 32, 33, 34, 35, 36, 37, 45

Grad-CAM Gradient-weighted Class Activation Mapping iv, xi, xii, 1, 2, 3, 4, 5, 7, 8, 12, 13, 14, 18, 19, 20, 21, 22, 25, 28, 29, 30, 32, 33, 34, 35, 36, 37, 45

IDG Idealized Domain Generalization . . . 53

ML Machine Learning . . . 4

PCE Polynomial Chaos Expansion . . . 9, 10

RDACN The Residual Dense Asymmetric Convolutional Neural Network 14, 15, 17, 25, 27, 28, 45

ReLU Rectified Linear Unit . . . xi, 19, 23, 24

SagNet Style Agnostic Networks . . . 54

SAR Synthetic Aperture Radar . . . 7, 8

SHAP SHapley Additive exPlanations . . . 11, 12

VH Vertical-Horizontal . . . 8

VV Vertical-Vertical 8

XAI Explainable Artificial Intelligence . . 1, 3, 4, 5, 7, 8, 10, 11, 12, 25, 27, 28, 46, 51, 55, 58

1. INTRODUCTION

In many remote sensing image analysis and classification-related tasks, the usage of deep-learning based approaches dominates the research domain. Despite their prevalence, most of the deep learning-based approaches lack explainability (Castelvecchi, 2016). The limited or absent interpretability of these models raises significant concerns regarding the reliability, reproducibility, and trustworthiness of their decisions and their consequences, especially in practical scenarios like banking or military applications. The lack of explanations for their decisions makes these models non-accountable. Therefore, this study delves into exploring methodologies aimed at shedding light on the decision processes of a deep learning model.

Explaining the internal decision-making mechanisms of deep neural networks is crucial, and explainable artificial intelligence (XAI) aims to achieve this objective. The absence of explanations regarding how deep neural networks work during predictions compromises the reliability and efficiency of the models. Knowing which features are more significant allows for the improvement of model efficiency by eliminating the less important ones. This not only enhances the transparency of the decision-making process but also utilizes the computational resources effectively. Furthermore, the element of randomness and unpredictability in the models' behavior cannot be effectively addressed without the explanations provided by XAI. As a result, this limitation constrains our ability to improve and refine the models' performance. Within this study, the interpretation tools were Grad-CAM (Selvaraju, Cogswell, Das, Vedantam, Parikh & Batra, 2017), Grad-CAM++ (Chattopadhay, Sarkar, Howlader & Balasubramanian, 2018), and Guided Backpropagation (Springenberg, Dosovitskiy, Brox & Riedmiller, 2015). These tools were integrated into a Residual Dense Asymmetric Convolutional Neural Network (RDACN) to understand where the model concentrates on during the prediction with respect to the specified category. Interpretation of the model's attention comes from the heatmaps generated from the last block of the model for Grad-CAM and Grad-CAM++. On the other hand, the Guided Backpropagation will highlight the regions in the input patch where the influence of the target category prevails. The details regarding the explainable artificial intelligence methods and their integrations are covered in chapter 3 and chapter 4.

In the subsequent chapters, three main experiments are covered in detail, revealing distinctions in the convolutional neural network's decision-making processes. The first experiment is based on comparing the impact of varying patch sizes, visualizing their influence on heatmap generation and spatial attention. The second experiment compares the Grad-CAM and Grad-CAM++ methods to provide insights regarding their limitations, as both methods generate heatmaps for the input image patches with respect to particular classes. It sheds light on their effectiveness in highlighting crucial regions during predictions across patch sizes of 7×7, 11×11, 15×15, 19×19, and 23×23 pixels. Lastly, the third experiment focuses on the outcomes of Guided Backpropagation, conducted with a fixed patch size per class. By combining all of these experiments, we gain insights into the dynamics of the model's attention concerning specific target classes and its decision processes across different scenarios.

In the initial experiment, the observation was that regardless of the patch size, the network directed its focus to the central part of the patch, as indicated by the generated heatmaps. This tendency was particularly noticeable with Grad-CAM++, except for the "Asphalt" and "Self-blocking Bricks" classes, where their heatmaps displayed diverse and distinct highlighted regions. This suggests that, for certain classes, attention of the convolutional neural network may not be consistently focused on the center during the prediction phase. In the second experiment, a comparison of heatmaps generated by Grad-CAM and Grad-CAM++ revealed that the heatmaps generated by Grad-CAM++ demonstrated heightened levels of activation in contrast to Grad-CAM. Notably, for a patch size of 11, Grad-CAM was unable to generate the heatmaps for "Gravel" and "Shadows" classes due to factors such as class imbalance within those classes, an insufficient patch size to capture variations in these classes and the complexity of the model. Concluding with the last experiment, the results demonstrated that the mean intensity levels per class were consistent through the spectral bands except for some classes where random spikes were observed. Moreover, the findings provided insights into the variation in patterns of each individual band that contributed to the model's predictions. These insights suggest that in future research, identifying specific bands makes it reasonable to eliminate less important spectral bands, thereby benefiting processing time.

Overall this study contributes to opening up the inherent black box nature of the deep neural networks by providing insights into their prediction-making phase with respect to specific classes, thereby, enhancing our interpretation in the under-explored domain of hyperspectral remote sensing image classification.

2. RELATED WORK

In this chapter, the background includes the challenges and advancements in the field of explainable artificial intelligence and its applications in remote sensing, particularly focusing on hyperspectral image analysis and classification. The primary sections cover related works about three common XAI methods found in the literature: Grad-CAM, Grad-CAM++, and Guided Backpropagation. The subsequent chapter introduces additional XAI studies within the remote sensing field. Finally, the last chapter discusses the existing literature, highlights accomplishments, identifies gaps, and introduces a summary of the methodologies employed in this thesis to fill these gaps.

2.1 Explainable Artificial Intelligence Techniques

Deep learning-based approaches constitute the state-of-the-art in most if not all, remote sensing image analysis-related tasks. However, their lack of interpretability and explainability, due to their "black box nature" has been a source of criticism ever since the inception of neural network (Castelvecchi, 2016). The term black box problem implies the difficulty that lies in the challenging nature of understanding how AI systems, predominantly neural networks, make their decisions or predictions. Current research is mainly based on unraveling this "black box" to gain insights into their inner mechanisms. It is imperative to have robust and intelligent deep learning models to prevent AI systems from being misled or biased towards specific decisions. This interpretability concern has been further emphasized in the context of Earth observation, where physical interpretations are of crucial significance.

In response to the limitations posed by the lack of interpretability in deep learning models, recent years have witnessed a growing body of research focused on developing explainable artificial intelligence (XAI) methods tailored for remote sensing applications (Gevaert, 2022; Kaya, Aptoula & Ertürk, 2023). Although complex deep learning models have high accuracies in this domain, their inherent lack of explainability nature remains a significant weakness (Gevaert, 2022). This study has comprehensively reviewed explainable ML and AI studies within Earth Observation, drawing distinctions between intrinsic versus post-hoc approaches, model-specific versus model-agnostic characteristics, and global versus local explanation methods. The paper has emphasized the need for social regularization for transparency and accountability for responsible AI models. Furthermore, the study by Kaya et al. (2023) supports the notion that the trajectory of XAI in remote sensing continues to grow. While these efforts have made significant strides, the complexities inherent in remote sensing data, especially in the context of hyperspectral images, present unique challenges that demand nuanced interpretability solutions.

Unlike traditional computer vision research areas, remote sensing commonly deals with images possessing multiple spectral bands and often containing relatively small (w.r.t. the image size) objects of interest. Applying existing XAI methods to this area, especially hyperspectral images with hundreds of bands at an often low spatial resolution, is not straightforward, though direly needed.

In this thesis, three main XAI methods are employed: Grad-CAM, Grad-CAM++, and Guided Backpropagation, to enhance the interpretability of the deep neural network used in the experiments covered in chapter 4. Gradient-weighted Class Activation Mapping (Grad-CAM), a widely used method, visualizes the significant regions of an image that have the most impact on the final model prediction. It generates heatmaps showing class significance by calculating the gradients of the target class scores with respect to the feature maps of the last convolutional layer.

Grad-CAM++ is built on Grad-CAM and improves the generated heatmaps by introducing a weighted combination of positive gradients. Additionally, the resulting heatmaps have improved precision in localizing crucial features and are better for detecting multiple objects belonging to the same class.

On the other hand, Guided Backpropagation highlights influential pixels through backpropagation of positive gradients to provide insights regarding the model's decision-making process. This method does not create heatmaps; instead, it highlights the crucial regions on the input image. These three methods shed light on the decision-making processes of deep neural networks, and the following sections provide related work regarding them.

2.2 Gradient-weighted Class Activation Mapping (Grad-CAM)

To date, only a handful of studies have ventured into the application of XAI methods in the field of remote sensing, and even fewer have tackled hyperspectral imagery. For instance, Gradient-weighted Class Activation Mapping (Grad-CAM) has recently been extended to 3D convolutional neural networks (CNNs) to address hyperspectral remote sensing image classification, with a specific focus on deployment in edge computing environments (De Lucia, Lapegna & Romano, 2022). The authors have employed spectral accumulation, where a single value per pixel represents the activation for the selected class on all spectral bands. This expansion makes it possible to visualize the activation volumes that are extracted from the neural network layers. Their proposed technique generates activation maps that reveal the significant choices made by the CNN within different layers by analyzing the final hidden layer preceding the fully connected layer in the models (De Lucia et al., 2022). Adapting Grad-CAM to pixel-wise spatio-spectral classifiers using 3D CNNs enables the interpretation of network decisions for hyperspectral remote sensing image classification, specifically in Edge Computing environments. The detailed calculation for this study is in section 3.2.

For investigating the behavior of deep neural network models when exposed to noisy input data, Gawlikowski, Ebel, Schmitt & Zhu (2022) employed Grad-CAM to generate class saliency maps. They found that the noisy cloud data caused the model to concentrate on the noise rather than the expected focus on land cover, implying a need for developing more robust models in the future.

One of the limitations of the Grad-CAM method is its inability to detect multiple objects in images, it is particularly evident in remote sensing image patches (Huang et al., 2022). To address this problem, Huang et al. (2022) developed a model named encoder-classifier-reconstruction CAM (ECR-CAM) neural network, comprising four modules: an encoder module, a classifier module, a reconstruction module, and a CAM module which can be seen in Figure 2.1. Extracting the image features is employed by the encoder module. Then, the reconstruction module, a crucial component, is primarily used to identify additional target objects. In order to reconstruct input images, involving a pixel-level process, the reconstruction module utilizes the extracted features. Throughout the reconstruction process, features are allowed to preserve vital information about all objects that cannot be achieved through the classification task alone. The last component, CAM, is utilized to reveal more target objects containing more informative features. The authors also demonstrated that their proposed model has enhanced classification performance and accurate localization of target objects.

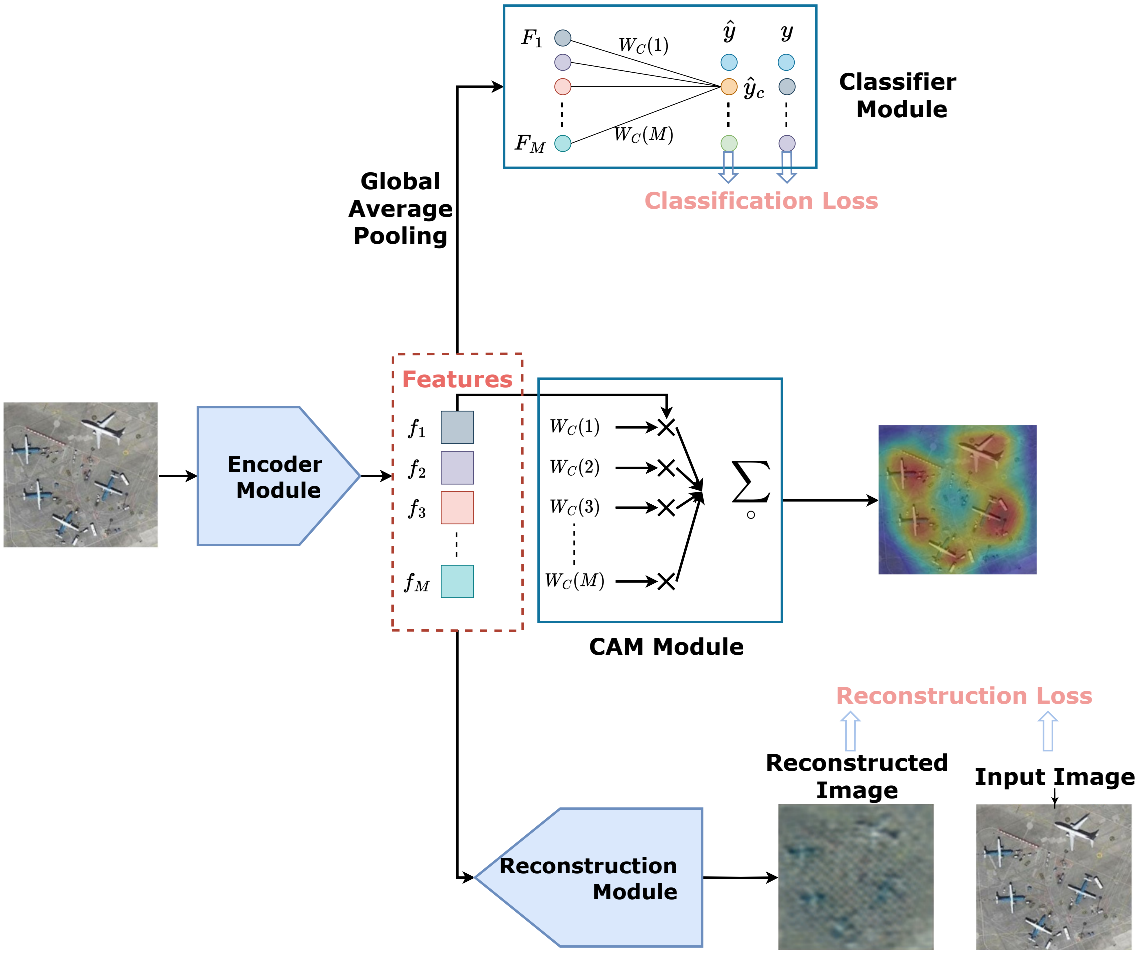


Figure 2.1 ECR-CAM Framework. An encoder module, a classifier module, a reconstruction module, and a CAM module are depicted. Encoder generates the image features $f = \{f_1, f_2, \ldots, f_M\}$ and they are transmitted to the remaining three modules. w_c denotes the representation vector for the c-th class, with $w_c(j)$ representing its j-th element (Huang et al., 2022).

2.3 Generalized Gradient-weighted Class Activation Mapping (Grad-CAM++)

Typically, Grad-CAM++ is commonly used to visualize the performance of models, with Tong, Chen, Han, Li & Wang (2020) providing an example in the remote sensing field. In their study, the authors used Grad-CAM++ to show their proposed channel attention network extracting important information from remote sensing images.

In another study, Hacıefendioğlu, Demir & Başağa (2021) utilized Grad-CAM++, Score-CAM (Wang, Wang, Du, Yang, Zhang, Ding, Mardziel & Hu, 2020) (which does not use gradients), and Grad-CAM to visualize the location of landslides in Rize, Turkey. Furthermore, Cai, Huang, He, Li, Qi, Peng, Zhou & Zhang (2023) used 1D-Grad-CAM++ to visualize and detect the crucial wavelengths that influence the performance of the model in the hyperspectral imaging of Radix Paeoniae Alba, a plant that is crucial in China.

In a novel approach, the study by Carneiro, Pádua, Peres, Morais, Sousa & Cunha (2022) employs segmentation as a preprocessing step for automatic image classification, particularly focusing on grapevines. They cleared the background from grapevine images to enhance classification accuracy. Furthermore, the authors applied Grad-CAM and Grad-CAM++ to illustrate the impact of segmentation on model predictions. It was observed that background pixels had less relevance for the model in making predictions.

For SAR data classification task, particularly in an automatic target recognition, Panati, Wagner & Brüggenwirth (2022) used Grad-CAM as one of the interpretative tools to visualize their proposed deep neural network architecture makes predictions based on target information.

Su, Zhang, Xiao, Li & Wang (2022) evaluated variations of CAM techniques, including Grad-CAM and Grad-CAM++ in the geographic object extraction task for a comparison study. According to their results, CAM was the most efficient method, Grad-CAM was the most accurate, and Grad-CAM++ achieved the highest integrity for this task. On the other hand, in another comparison study by Kakogeorgiou & Karantzalos (2021) on multi-label deep learning classification, it was stated that no XAI method emerged as the superior choice. The authors concluded that Grad-CAM was the computationally efficient method compared to other methods.

In addition to the previously mentioned generic Grad-CAM++ method, Gao, Liu, Li, Hou, Li & Zhao (2023) introduced a method for generating saliency maps referred to as augmented high-order gradient weighting class activation mapping (augmented Grad-CAM++). Their proposed method utilizes image geometry augmentation and super-resolution techniques for improving the accuracy of target localization and produce higher resolution saliency maps, to address the limitations of existing visual interpretation techniques in the deep learning literature. Instead of using a single input image, they create a set of augmented images from the input image and then produce the activation mappings individually. In the next step, they generate the resulting saliency map by combining the augmented activation mappings. In the last stage, their super-resolution method is applied to create higher-resolution saliency maps by adding pixel points for reconstructing the saliency map pixels.

2.4 Guided Backpropagation

In their work, Su, Cui, Guo, Zhang & Yu (2022) employed Guided Backpropagation method (with other XAI techniques including Grad-CAM) in synthetic aperture radar (SAR) image classification. They presented their insights into model decisions for SAR. They discovered that Guided Backpropagation is among the lowest scores in VH-polarization and VV-polarization parts of OpenSARUrban dataset (Zhao, Zhang, Yao, Datcu, Xiong & Yu, 2020).

For the task of SAR target recognition, Xu, Sun, Chen, Lei, Ji & Kuang (2021) investigated the robustness of deep neural networks. They assessed the risk posed by adversarial examples and then integrated adversarial contrastive pretraining into SAR target recognition, introducing their unsupervised defense approach. To visually interpret the effectiveness of their defense method, they used Guided Backpropagation and observed that adversarial examples harm the activation of the model, causing the standard model to focus on the whole region. When they applied their proposed defense method, their model primarily focused on the core regions in the images, thereby enhancing the adversarial robustness of their model.

Another application is in hyperspectral image processing, particularly the study of band selection by Zhao, Zeng, Liu & He (2020). Selection of the bands was based on XAI using an improved version of Grad-CAM to generate gradient-weighted heatmaps (GradHM). Additionally, they applied Guided Backpropagation to further enhance the details of the heatmaps and named the resulting heatmaps as Guided-GradHM. They proposed two methods: Average Selection (AS) and Total Selection (TS), to combine information coming from GradHM and Guided-GradHM, resulting in four combinations: GradHM+AS, GradHM+TS, Guided-GradHM+AS, and Guided-GradHM+TS.

The average of the heatmaps is calculated by taking the average heatmap for all categories. The average selection (AS) method selects bands from the top N/l ranked values for each category, excluding bands already chosen, where N is the number of selected bands from the full bands, and l is the number of categories. The total selection (TS) method creates a total heatmap composed of average heatmaps for all categories, from which N bands are selected based on the top N-ranked values.

They found that Guided-GradHM is helpful in highlighting the fine-grained details in the bands, hence mitigating the side effects of adjacent band correlation.

Zhang, Zhao & Li (2021) also employed Guided Backpropagation to visualize features extracted by transformers. For spatial-variant convolutional neural networks (SV-CNN) proposed by Dai, Jin, Song, Sun & Wu (2020), Guided Backpropagation is employed to identify significant information in input samples and show position-coding significance on each layer. The authors' findings indicated that information through position-coding is crucial for the SV-CNN.

2.5 Other XAI Techniques

One of the notable approaches involves a global model distillation approach which is a method proposed to replace black box models with fully explainable surrogate models utilizing polynomial chaos expansion (PCE) (Taskin, 2022). Utilization of PCE is done to distill the essential insights from complex models, offering a more transparent understanding of the decision-making process. Their proposed method consists of two stages. In the first stage, they generated random samples within the input space and fed them into a pre-trained black box model, such as a random forest, to have corresponding outputs of the model. This first step guarantees to capture the essential nature of the black box model's behavior.

Subsequently, the dataset produced in the initial phase is used to build a surrogate model through PCE regression, with a focus on two hyperspectral datasets: Botswana and Salinas Valley. A probability function is then customized for each feature, which guides the generation of random inputs for sampling. The reason for using PCE regression to create a surrogate model is motivated by the aim to obtain a mapping from input samples to the output variable. In other words, PCE reveals the functional connection between inputs and outputs. Then the estimation of the coefficients of these polynomials is achieved through least-squares optimization.

The performance of the surrogate model is evaluated by comparing its accuracy with the original black box model, showing its capability to successfully replace the complex nonlinear model in hyperspectral image classification.

The following is the formulation of how the mapping from random input samples

$X \in R^n$ with a given joint probability density function is calculated:

$$Y = \sum_{\alpha \in R^n} y_\alpha \Phi_\alpha(X) \tag{2.1}$$

y_α represents coefficients that are going to be computed and $\Phi_\alpha(X)$ represents the basis function which is composed of multivariate orthonormal polynomials:

$$\Phi_\alpha(x) = \prod_{i=1}^{n} \Phi_{\alpha_i}^{(i)}(x_i) \tag{2.2}$$

In summary, the mentioned method utilizes PCE as a surrogate model to replace a black box model in hyperspectral image classification. When the surrogate model is built using PCE regression, it provides a more interpretable and straightforward alternative to the black box model.

For the hyperspectral pixel classification task, researchers have explored novel solutions such as the conversion of hyperspectral pixels into spectral graphs, followed by convolution (Deshpande, Thakur & Balamuralidhar, 2021). This study aims to improve the interpretability of spectral features and their contributions to pixel classification tasks. The researchers achieved this objective by generating an n-dimensional spectral graph representation utilizing it as an input for their CNN. They decomposed the hyperspectral shape through hierarchical features coming from various convolution layers and levels. Additionally, their filters were able to learn edges, arcs, arc segments, and similar shape features. This study offered two networks, one based on a one-dimensional architecture inspired by Dai, Dai, Qu, Li & Das (2016), and the other inspired by CapsuleNet (Sabour, Frosst & Hinton, 2017). In addition to them, an intrinsically interpretable method has been used by Wang, Abliz, Ma, Liu, Kurban, Halik, Pietikäinen & Wang (2022) for soil copper concentration estimation based on an attentive interpretable tabular learning model (TabNet) which has also a high accuracy. It aims to reduce data processing time and select features based on sequential attention. Since the selection of each feature at every step is understandable it makes it an interpretable model. Its feature selection mechanism is based on a mask matrix which ensures the chosen features are sparse and non-repetition in them. By using a combination of a decision tree and a feature transformer layer, their model is able to process and integrate features efficiently.

The study by Gizzini, Shukor & Ghandour (2023) investigated a less explored area, namely, image segmentation. It is one of the recent works that adapted CAM-based XAI techniques to remote sensing image segmentation. Additionally, they proposed an XAI metric for evaluating model uncertainty based on entropy in the segmentation of target class pixels.

Although pixel classification is commonly handled via semantic segmentation techniques in computer vision (Csurka, Volpi & Chidlovskii, 2023), this requires the availability of fully labeled (at pixel-level) datasets. The scarcity of remote sensing experts and hyperspectral datasets, along with the sparsity of labels in benchmark hyperspectral classification datasets (such as Pavia University, Indian Pines, etc.), prohibit the use of semantic segmentation in this context. Consequently, the vast majority of the hyperspectral remote sensing classification state-of-the-art resorts to patch-based spectral-spatial strategies (Li, Song, Fang, Chen, Ghamisi & Benediktsson, 2019), centered on the pixel of interest. This widely encountered approach relies on the implicit assumption that the network models focus on the central pixel and its relation to its spatial surroundings during training.

Another technique that is designed to investigate the role of each input feature in the predictions generated by a machine learning model is called SHapley additive exPlanations (SHAP) (Lundberg & Lee, 2017). It originates from Shapley value (Shapley, 1953) in game theory that equally distributes payoffs in a game among all participating players. When adapted to machine learning, the predictions made by the model refer to the game, the input features influencing those predictions denote the players and the contribution of each feature to the overall prediction corresponds to the payoff. To generate an interpretable additive approximate model, denoted as $f(x)$, for a complex original machine learning model, $g(x)$, SHAP can be calculated as follows, where ϕ_0 represents the bias term, and ϕ_i denotes individual feature contributions:

$$f(x) = \phi_0 + \sum_{i=1}^{n} \phi_i x_i \tag{2.3}$$

In general, the prediction $f(x)$ is decomposed into a summation of contributions from individual features, offering distinct importance scores for each feature while preserving the intricate relationships among the features. The Shapley values consider both independent and combined influences of variables, with positive values indicating that the corresponding features improve the prediction and negative values indicating a decrease in the prediction. The magnitude shows the feature's impact on the prediction. Additionally, since SHAP takes into account all possible cases for features, it can be considered both a global and local XAI method, with a thorough understanding of the model as a whole and detailed insights into the effects of individual features to the model's prediction (Kaya et al., 2023).

In the context of image classification, SHAP is a common method employed to discover the complexities of feature importance. Sahin, Erturk & Aptoula (2023) employed this method for hyperspectral image classification. The authors aimed to extract interpretations regarding spectral bands in the classification task, leveraging the Pavia University dataset.

In order to calculate the band-based mean SHAP values per class, they applied treeSHAP (Lundberg, Erion & Lee, 2018) to the Random Forest classifier trained on the Pavia University dataset. The impacts of each band's influence on nine classes are illustrated in their work. Furthermore, they visualized the twenty most crucial spectral bands, exhibiting the highest SHAP values along with their corresponding importance for each class.

2.6 Discussion

When examining the current literature, it becomes evident that there is an opportunity for further development in the field of XAI in hyperspectral image classification. As indicated by the aforementioned studies, many interpretative tools have been utilized to visualize model predictions in image classification. However, the exploration of such tools in the context of hyperspectral imaging is relatively limited. It is worth noting that the study conducted by Sahin et al. (2023) closely aligns with our focus by exploring band selection in the Pavia University dataset. Nevertheless, it employs SHAP, which does not directly consider the visual significance, particularly in terms of highlighting fine-grained features. As can be seen in their work, they considered the significance of bands with respect to SHAP and observed variations in bands for different categories. The variations of bands with respect to certain classes show parallelism with our findings in subsection 4.4.3.

Furthermore, since interpretation techniques such as Grad-CAM, Grad-CAM++, and Guided Backpropagation provide insights into where the model focuses on the input images, these interpretations would be valuable for users, especially in environmental monitoring, and similar fields where confidence in decisions plays a crucial role. Additionally, since these tools detect significant regions in the model predictions, unexpectedly highlighted regions in generated heatmaps might indicate issues with the training data or model parameters, requiring fine-tuning of the model or improvement of the training data. Understanding which parts of input image patches are crucial for model predictions is advantageous in resource allocation, creating room for the elimination of redundant parts.

To the best of our knowledge, and based on the literature reviewed in this thesis, no study has examined the impact of different patch sizes on heatmap generation using Grad-CAM and Grad-CAM++, as well as the influence of spectral bands on model predictions using Guided Backpropagation techniques in the domain of hyperspectral images within the Pavia University dataset. Therefore, in response to these aforementioned gaps in the existing literature, we addressed these issues in this thesis.

3. METHODS

This chapter begins by introducing the Residual Dense Asymmetric Convolutional Network (RDACN) as the central model for hyperspectral image classification, which is a relatively new and effective model in addressing challenges like low feature discrimination and a redundant number of network parameters. Moreover, RDACN's components, including its unique features such as asymmetric convolution and residual dense asymmetric convolutional blocks, are outlined. The following sections explain visualization techniques like Grad-CAM, Grad-CAM++, and Guided Backpropagation. These structures and methodologies collectively form the baseline for subsequent experiments while providing both enhanced classification capabilities and interpretability in the context of hyperspectral image classification.

3.1 RDACN Model

As can be seen from the previous chapter, the CNNs have strong performance in this domain. However, they often suffer from low feature discrimination and have an excessive amount of network parameters. The Residual Dense Asymmetric Convolutional Neural Network (Meng, Zhang, Zhao, Liu & Chang, 2022) is specifically designed to address previous problems and has been selected as the backbone CNN model in this thesis. This convolutional neural network model has been specifically designed for hyperspectral image classification in order to deal with low-level feature discrimination and the high number of network parameters and employs residual and dense connections within the network to improve classification accuracy and improve the previous layers' information. In order to collect hyperspectral features, two feature fusion methods of addition and channel stacking have been introduced. Additionally, the ordinary square convolutional kernel has been replaced with the asymmetric convolutional kernels, resulting in a reduced number of CNN parameters compared to the state-of-the-art.

RDACN comprises three main components: two residual dense asymmetric convolution blocks and a transition layer. As outlined by Meng et al. (2022) and in this thesis, the Pavia University hyperspectral image is initially partitioned into 3D cubes. Subsequently, a 3×3 convolution layer is applied to extract primitive features from those cubes. In the next step, distinctive spectral-spatial features are captured by residual dense asymmetric convolution blocks. The concatenation logic following the 1×1 convolution is effective in minimizing feature redundancy, while the asymmetric convolutional layers play a crucial role in reducing the number of parameters in RDACN compared to conventional convolutional layers. Next, the dimension of the features is modified by the transition layer which has 1×1 convolution that increases the number of output channels. In the final step, the linear classifier layer is employed to generate the final classification map. The overall architecture of RDACN can be seen in Figure 3.1

3.1.1 Asymmetric Convolution

In its nature, the $k \times k$ convolution operation can be replicated using asymmetrical convolutional layers with $1 \times k$ and $k \times 1$ kernels (Meng et al., 2022). This modification, as demonstrated by Ding, Guo, Ding & Han (2019), has been shown to enhance efficiency by augmenting the kernel structure of CNNs and reducing the number of parameters. Consequently, in the residual dense asymmetric convolutional block, Meng et al. (2022) replaced the 3×3 kernel with 1×3 and 3×1 kernels. This modification not only reduced the number of parameters in the network but also enriched the expression of features.

3.1.2 Residual Dense Asymmetric Convolutional Block

One of the advantages of a residual block is its ability to reutilize input features with the help of residual connection so that it can be helpful to improve classification accuracy. When compared to the feature fusion approach having addition in the residual block, the densely connected block directly concatenates the output features of the convolutional layers along the channel dimension (Meng et al., 2022). This approach enables the comprehensive utilization of information across all convolutional layers.

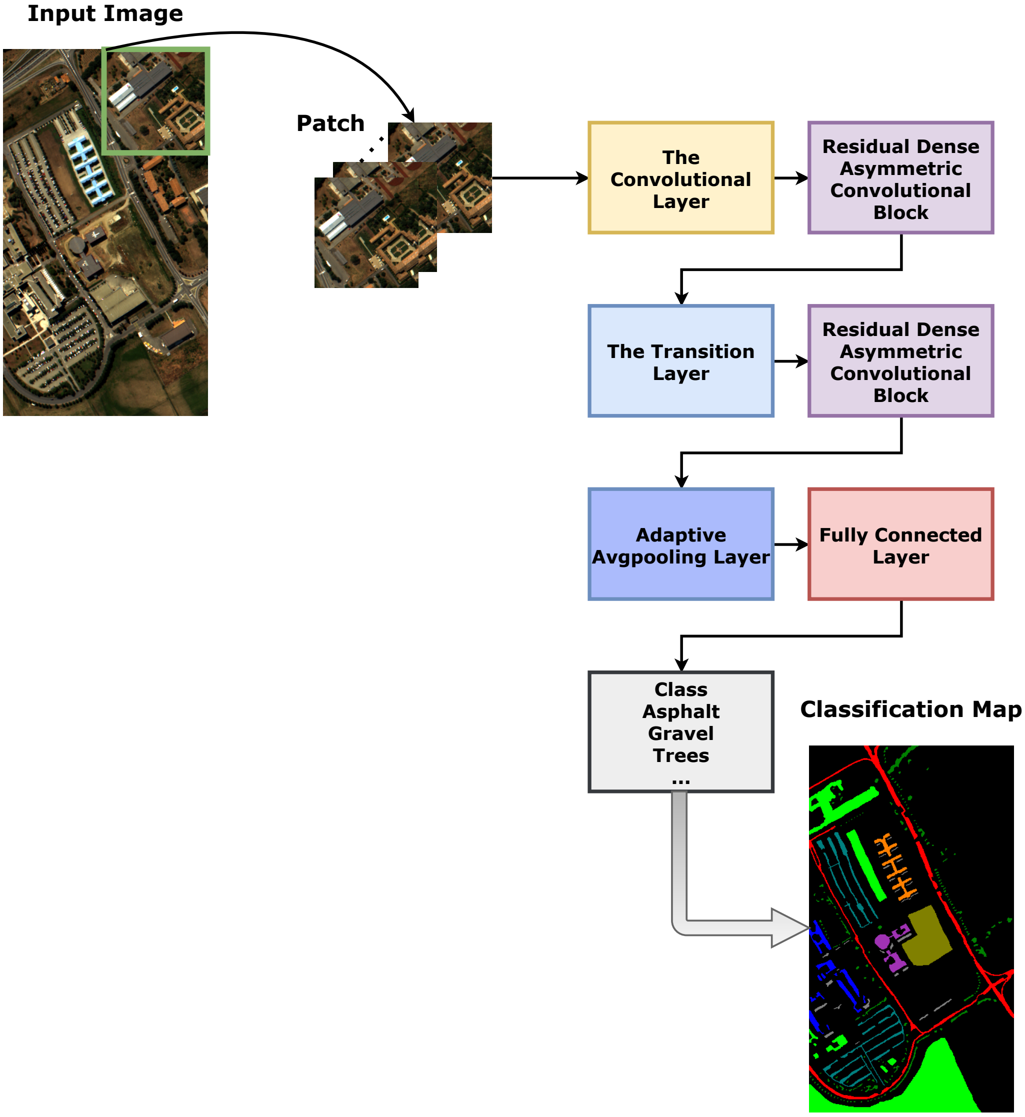


Figure 3.1 Residual Dense Asymmetric Convolutional Network Framework

In order to reduce a large number of redundant parameters due to having too many dense connections, Meng et al. (2022) proposed an innovative concatenation method. They only used dense connections with the input feature and the output features of the initial convolutional layer, differing from the conventional dense connection method where there are dense connections across multiple layers. Moreover, the resulting features from the 1×1 convolutional layer are combined at the left and right ends of the input, which are used as the input feature for the following convolutional layer. Furthermore, the summation of input and the last convolutional layer's final outputs is done via the residual connection to be able to reuse the features. They also changed the 3×3 convolutional layer to 1×3 followed by a 3×1 asymmetric convolutional layer in order to have fewer parameters in the residual dense asymmetric convolution block which is illustrated in Figure 3.2. This thesis concentrates on the last asymmetric block layer of RDACN.

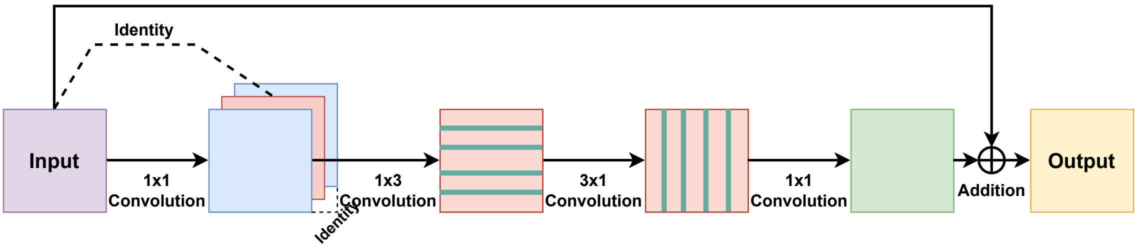


Figure 3.2 Residual Dense Asymmetric Convolutional Block

3.2 GradCAM

Before mentioning the details directly related to the remote sensing explainable artificial intelligence, it is crucial to mention Class Activation Maps (CAM). They are used in visualizing the important regions in the deep learning models which is tremendously popular in computer vision and deep learning. CAMs actually highlights the image regions that are important for the employed CNN for a specific class.

The logic behind CAM lies in the concept of utilizing the global average pooling layer in a CNN. The weights are calculated for every feature map at the last convolutional layer based on their significance in classifying into a specific category. Determining the weights is a derivation from the fully connected layer tied to the predicted category.

Taking the weighted combination of feature maps results in a heatmap that highlights the important regions in an input image, essential for the model to make accurate predictions. Additionally, the discriminative regions are localized after overlaying that heatmap onto the input image.

To explain the mechanism for CAM mathematically, for an input image, $f_k(x,y)$ denotes the activation of unit k in the final convolutional layer at spatial location (x,y) (Zhou, Khosla, Lapedriza, Oliva & Torralba, 2015). Subsequently, applying global average pooling for unit k results in $F^k = \sum_{x,y} f_k(x,y)$. Therefore, for a specific category c, input to the softmax is $S_c = \sum_k w_k^c F_k$ where w_k^c represents the weight related to category c for unit k. In essence, w_k^c signifies how important F_k

is for category c. P_c, the softmax output for category c is determined by $\frac{\exp(S_c)}{\sum_c \exp(S_c)}$ (Zhou et al., 2015). The bias term is omitted by the researchers and the input is explicitly set. With the substitution of $F_k = \sum_{x,y} f_k(x,y)$ into the class score, S_c, the result is as follows:

$$S_c = \sum_k w_k^c \sum_{x,y} f_k(x,y) = \sum_{x,y} \sum_k w_k^c f_k(x,y) \tag{3.1}$$

The class activation map for category c is denoted as M_c, with each spatial element being determined as follows:

$$M_c(x,y) = \sum_k w_k^c f_k(x,y). \tag{3.2}$$

Finally, they demonstrated that $S_c = \sum_{x,y} M_c(x,y)$, and hence $M_c(x,y)$ provides a direct implication of the significance of the activation at spatial coordinates (x,y) in classifying an image to category c.

With the intent of unveiling the opaque properties that are inherent in complex deep neural networks, it is crucial to understand the inner working mechanisms of their "black box" nature. Since the layers in the deep neural networks contain information regarding the input image features, it will be helpful to look at and extract information from them. One of the methods that are helpful in order to visually understand the hidden information in the layers is called Gradient-weighted Class Activation Mapping (Selvaraju et al., 2017). Similar to Class Activation Mapping, Grad-CAM generates a heat map with coarse-grained visualizations. In order to generate the heatmap, Grad-CAM feeds the gradients for a target class into the final convolutional layer and computes an importance score based on the gradients. The highlighted regions in the resulting coarse localization map refer to the crucial regions in the image for the classification (Selvaraju et al., 2017).

One of the significant advantages of Grad-CAM is that it can be applied to various CNN based models including those with fully-connected layers, employed for structured outputs, and those utilized in tasks with multi-modal inputs (Selvaraju et al., 2017). It can be utilized in the aforementioned tasks without the need for architectural modifications or re-training.

Features that are collected by convolutional filters, inherently preserve spatial information which the fully-connected layers might possibly lose. As a consequence, the last convolutional layers are expected to have an optimal balance between intricate spatial details and high-level information. Neurons in these layers are responsible for searching class-related information and Grad-CAM calculates the gradient for the significance of each neuron related to the objective. The information coming from the gradient is transmitted to the final convolutional layer.

$L^c_{\text{Grad-CAM}} \in \mathbb{R}^{u\times v}$ refers to the class-discriminative localization map of class c with width u and height v. Its first step is to calculate the gradient of the class score, $\frac{\partial y^c}{\partial A^k}$, where A^k represents the activation of feature maps, y^c is the class score before softmax (Selvaraju et al., 2017). Then, the neuron importance weights w^c_k are calculated as follows (Z stands for a constant value representing the total count of pixels in the activation map):

$$w^c_k = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A^k_{ij}}}_{\text{gradients via backprop}} \tag{3.3}$$

Afterward, Rectified Linear Unit (ReLU) is applied to the weighted combination of forward activation maps as follows:

$$L^c_{\text{Grad-CAM}} = \text{ReLU} \underbrace{\left(\sum_k w^c_k A^k\right)}_{\text{linear combination}} \tag{3.4}$$

The reason behind applying ReLU is to focus solely on features with a positive impact on the class of interest. In other words, only pixels with increasing intensity that would lead to the enhancement of y^c are taken into consideration.

It is worth mentioning that if an image contains multiple instances of the same object, Grad-CAM may not accurately highlight each. Additionally, the localization might include some parts of the object of interest rather than the entire object due to calculating the unweighted average of partial derivatives.

In the study of De Lucia et al. (2022), the authors have adapted the current formula for computing the class-discriminative localization map to apply the same logic for 3D CNNs. While the 2D Grad-CAM method calculates the localization maps for a 2D image with dimensions $u \times v$ pixels, where u represents image width and u represents image height, the extended version of 3D CNNs involves updating the equation to be applied to volumes different than 2D images.

To interpret the decisions of a certain convolution layer, y^c represents the score for class c before softmax. If the layer has k feature maps A^k with dimension $H \times W$, the neuron importance weights can be calculated using the following formula:

$$\alpha_k^c = \frac{1}{H \cdot W} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{3.5}$$

Here, $\frac{\partial y^c}{\partial A_{ij}^k}$ represents the pixel-wise gradients of y^c with respect to A^k by backpropagation, H is the spatial height, W is the spatial width, and D is the depth. The Grad-CAM class-discriminative localization map of the resolution $u \times v$ of the input image as a weighted combination is calculated as follows:

$$L_{\text{Grad-CAM}}^c = UP_{u,v}\left(\mathrm{Re}\,LU\left(\sum_k \alpha_k^c A^k\right)\right) \in \mathbb{R}^{u \times v} \tag{3.6}$$

$UP_{u,v}$ represents a bilinear upsampling function, the features that have negative contributions to class c are eliminated by the Rectified Linear Unit (ReLU) activation function.

By extending this formulation to volumes, we can obtain 3D Grad-CAM activation maps for 3D CNNs as $L_{\text{Grad-CAM}}^c \in \mathbb{R}^{u \times v \times w}$ where channel k has dimensions $H \times W \times D$. If A_{ijb} is the intensity of activation for a pixel p_{ij} in the $b \in B$ band for a specific class with pixel neighborhood of $(2N+1) \times (2N+1)$ dimension where N is the number of pixels, the following is the 3D Grad-CAM calculation:

$$A_{ijb} = \sum_{h=-N}^{N} \sum_{k=-N}^{N} A_{ijb}^{hk} \tag{3.7}$$

In order to calculate the 3D Grad-CAM activation maps, they have used `M3d-CAM` library by Gotkowski, González, Bucher & Mukhopadhyay (2020), which contains tools to obtain Grad-CAM activation maps for 3D CNNs in PyTorch library (Gotkowski et al., 2020). This study helps to obtain the activation maps by utilizing the last hidden layer of the model before the fully connected layer, which is usually considered as the most suitable choice for extracting and interpreting explanations about 3D CNNs. In summary, the mentioned study extends the traditional Grad-CAM method to 3D CNNs by renewing the equation used to calculate the 2D class-discriminative localization maps and using the `M3d-CAM` library for creating the activation maps. These activation maps are then used to visualize the activation volumes and make interpretations of the choices made within the neural network layers.

3.3 Grad-CAM++

Grad-CAM++ (Chattopadhay et al., 2018) is an improved version of Grad-CAM that further enhances the interpretability of the heatmaps by addressing the aforementioned inconveniences. It employs a weighted combination of positive gradients with respect to a specific class score in order to quantify the contributions toward the final class activation.

$$w_k^c = \sum_i \sum_j \alpha_{ij}^{kc} \operatorname{ReLU}\left(\frac{\partial y^c}{\partial A_{ij}^k}\right) \tag{3.8}$$

Once again, the Z term here refers to the constant value representing the total count of pixels in the activation map. The term $\frac{\partial y^c}{\partial A_{ij}^k}$ denotes the gradient of the class c score, y^c, with respect to the activation A_{ij}^k at the pixel position of (i, j) for the k^{th} feature map. The key difference lies in the additional summation in the calculation. The normalization and second summation over spatial dimensions are helpful for calculating each spatial location's importance in the feature maps. Hence, the final class saliency maps are calculated as:

$$L_c = \operatorname{ReLU}\left(\sum_k \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} A_{ij}^k\right) \tag{3.9}$$

$$Y^c = \sum_k \left[\sum_i \sum_j \left\{\sum_a \sum_b \alpha_{ab}^{kc} \cdot \operatorname{relu}\left(\frac{\partial Y^c}{\partial A_{ab}^k}\right)\right\} A_{ij}^k\right] \tag{3.10}$$

After taking the partial derivative of both sides of the equation with respect to A_{ij}^k, the equation becomes as follows:

$$\frac{\partial Y^c}{\partial A_{ij}^k} = \sum_a \sum_b \alpha_{ab}^{kc} \cdot \frac{\partial Y^c}{\partial A_{ab}^k} + \sum_a \sum_b A_{ab}^k \left\{\alpha_{ij}^{kc} \cdot \frac{\partial^2 Y^c}{\left(\partial A_{ij}^k\right)^2}\right\} \tag{3.11}$$

Then taking an another derivative with respect to A_{ij}^k results in:

$$\frac{\partial^2 Y^c}{\left(\partial A_{ij}^k\right)^2} = 2 \cdot \alpha_{ij}^{kc} \cdot \frac{\partial^2 Y^c}{\left(\partial A_{ij}^k\right)^2} + \sum_a \sum_b A_{ab}^k \left\{\alpha_{ij}^{kc} \cdot \frac{\partial^3 Y^c}{\left(\partial A_{ij}^k\right)^3}\right\} \tag{3.12}$$

After rearranging the terms, the resulting equation becomes as follows:

$$\alpha_{ij}^{kc} = \frac{\frac{\partial^2 Y^c}{\left(\partial A_{ij}^k\right)^2}}{2\frac{\partial^2 Y^c}{\left(\partial A_{ij}^k\right)^2} + \sum_a \sum_b A_{ab}^k \left\{ \frac{\partial^3 Y^c}{\left(\partial A_{ij}^k\right)^3} \right\}} \tag{3.13}$$

Thus, the localization accuracy improves by this modification, and the resulting heatmaps are more precise and visually informative, even in the case of multiple objects of the same class present in the input scene. Figure 3.3 visualizes the difference in calculations between CAM, Grad-CAM, and Grad-CAM++ techniques.

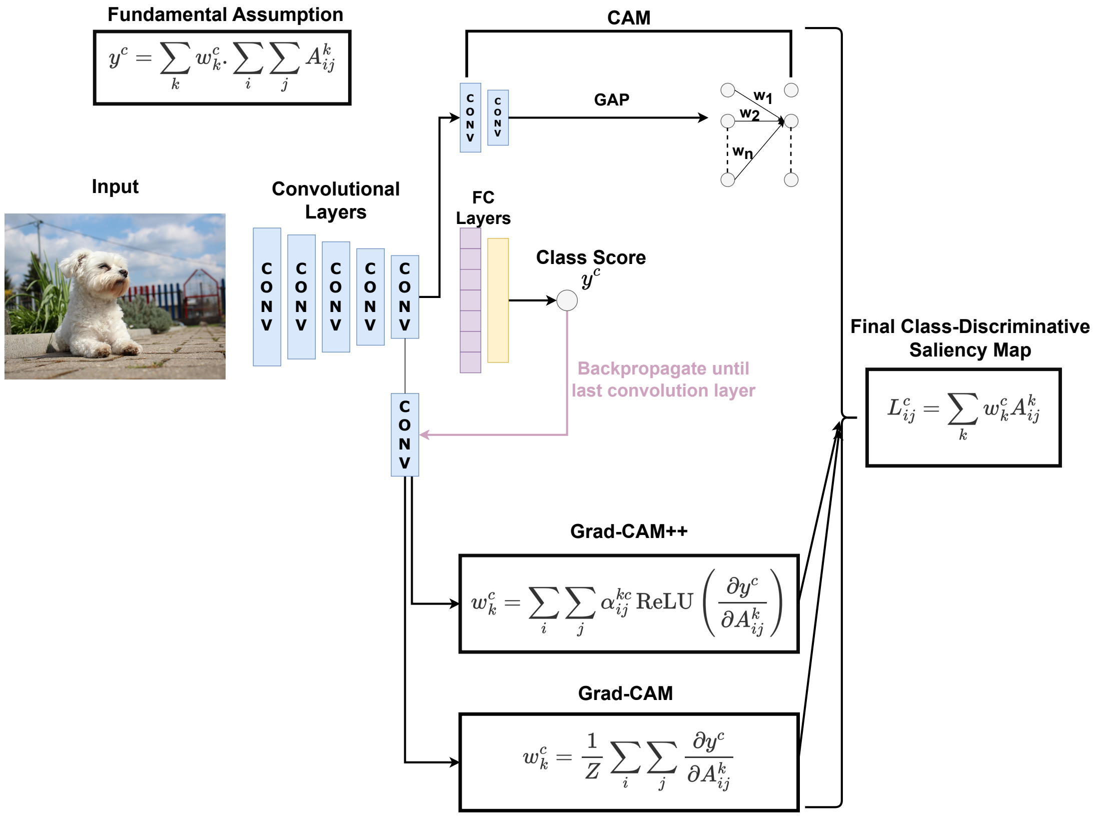

Fundamental Assumption
$y^c = \sum_k w_k^c . \sum_i \sum_j A_{ij}^k$
CAM
GAP
w1
w2
wn
Input
Convolutional Layers
CONV
FC Layers
Class Score
y^c
Backpropagate until last convolution layer
Final Class-Discriminative Saliency Map
$L_{ij}^c = \sum_k w_k^c A_{ij}^k$
Grad-CAM++
$w_k^c = \sum_i \sum_j \alpha_{ij}^{kc} \, \mathrm{ReLU}\left(\frac{\partial y^c}{\partial A_{ij}^k}\right)$
Grad-CAM
$w_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}$

Figure 3.3 Comparison of CAM, Grad-CAM, Grad-CAM++

3.4 Guided Backpropagation

Guided Backpropagation is used to visualize the importance of different image features for a neural network's final prediction (Springenberg et al., 2015). The gradients of the input image are calculated with respect to the network's output, while the negative gradients are masked out. It highlights the regions of the image that positively contribute to the activation of the target class by backpropagating only positive gradients. This way, it reveals the importance of different image regions that influence the network's decision. Since they are masking out values for which at least one of the values in the top gradient or bottom is negative it actually introduced an additional guidance signal from higher layers to the usual backpropagation while preventing the negative gradient's backward flow. Hence the name "Guided Backpropagation" comes from this additional signaling mechanism (the forward and backward pass can be seen in Figure 3.4).

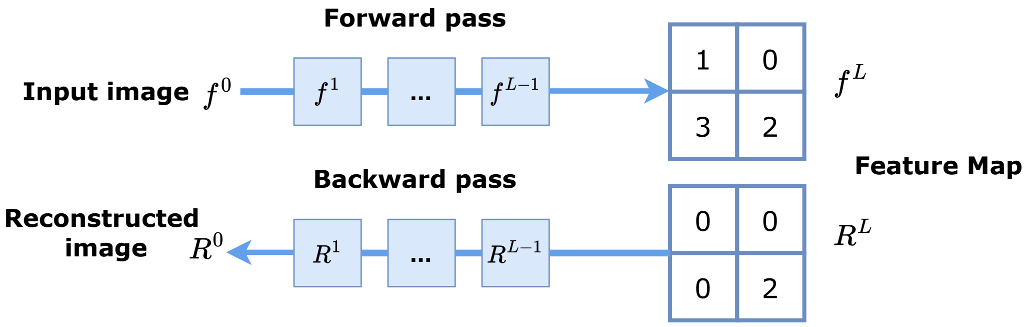


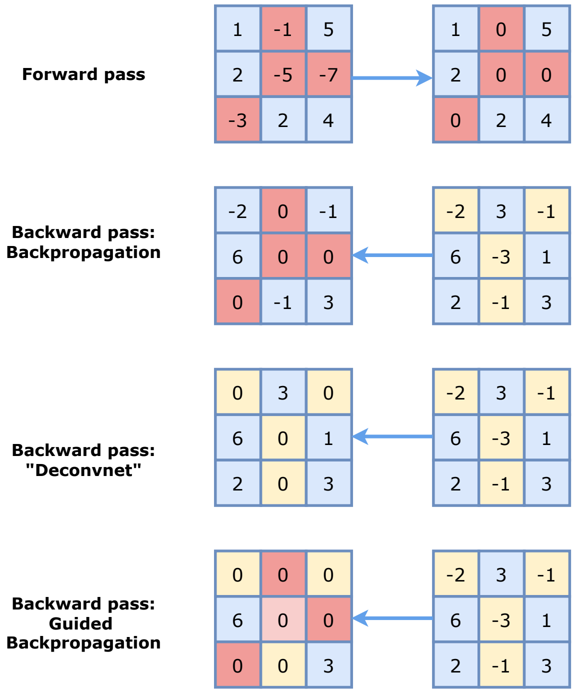
Figure 3.4 The visualization of higher layer neuron activations. Forward pass is performed to a layer when an input image is given. Then all of the activations are zeroed out except one, and this signal is propagated back to to obtain a reconstructed image R^0

The output of an activation function f for a ith neuron in layer $l+1$ can be calculated as follows:

$$\text{Activation:} \quad f_i^{l+1} = \text{relu}\left(f_i^l\right) = \max\left(f_i^l, 0\right) \tag{3.14}$$

Following is the common approach to backpropagate an output activation (out) through a ReLU function in layer l:

$$\text{Backpropagation:} \quad R_i^l = \left(f_i^l > 0\right) \cdot R_i^{l+1}, \text{ where } R_i^{l+1} = \frac{\partial f^{\text{out}}}{\partial f_i^{l+1}} \tag{3.15}$$



Forward pass

$$\begin{bmatrix}1&-1&5\\2&-5&-7\\-3&2&4\end{bmatrix} \rightarrow \begin{bmatrix}1&0&5\\2&0&0\\0&2&4\end{bmatrix}$$

Backward pass: Backpropagation

$$\begin{bmatrix}-2&0&-1\\6&0&0\\0&-1&3\end{bmatrix} \leftarrow \begin{bmatrix}-2&3&-1\\6&-3&1\\2&-1&3\end{bmatrix}$$

Backward pass: "Deconvnet"

$$\begin{bmatrix}0&3&0\\6&0&1\\2&0&3\end{bmatrix} \leftarrow \begin{bmatrix}-2&3&-1\\6&-3&1\\2&-1&3\end{bmatrix}$$

Backward pass: Guided Backpropagation

$$\begin{bmatrix}0&0&0\\6&0&0\\0&0&3\end{bmatrix} \leftarrow \begin{bmatrix}-2&3&-1\\6&-3&1\\2&-1&3\end{bmatrix}$$

Figure 3.5 Comparative applications of backpropagation techniques through nonlinearity of ReLU

The deconvolutional network (deconvnet) calculation is as follows:

$$\text{Backward `deconvnet':}\quad R_i^l = \left(R_i^{l+1} > 0\right) \cdot R_i^{l+1} \tag{3.16}$$

As can be seen from the following equation, only the positive output and reconstructed image values are backpropagated different from regular and deconvnet backpropagation.

$$\text{Guided Backpropagation:}\quad R_i^l = \left(f_i^l > 0\right) \cdot \left(R_i^{l+1} > 0\right) \cdot R_i^{l+1} \tag{3.17}$$

The Figure 3.5 illustrates the comparison of different backpropagation techniques through the nonlinear ReLU function.

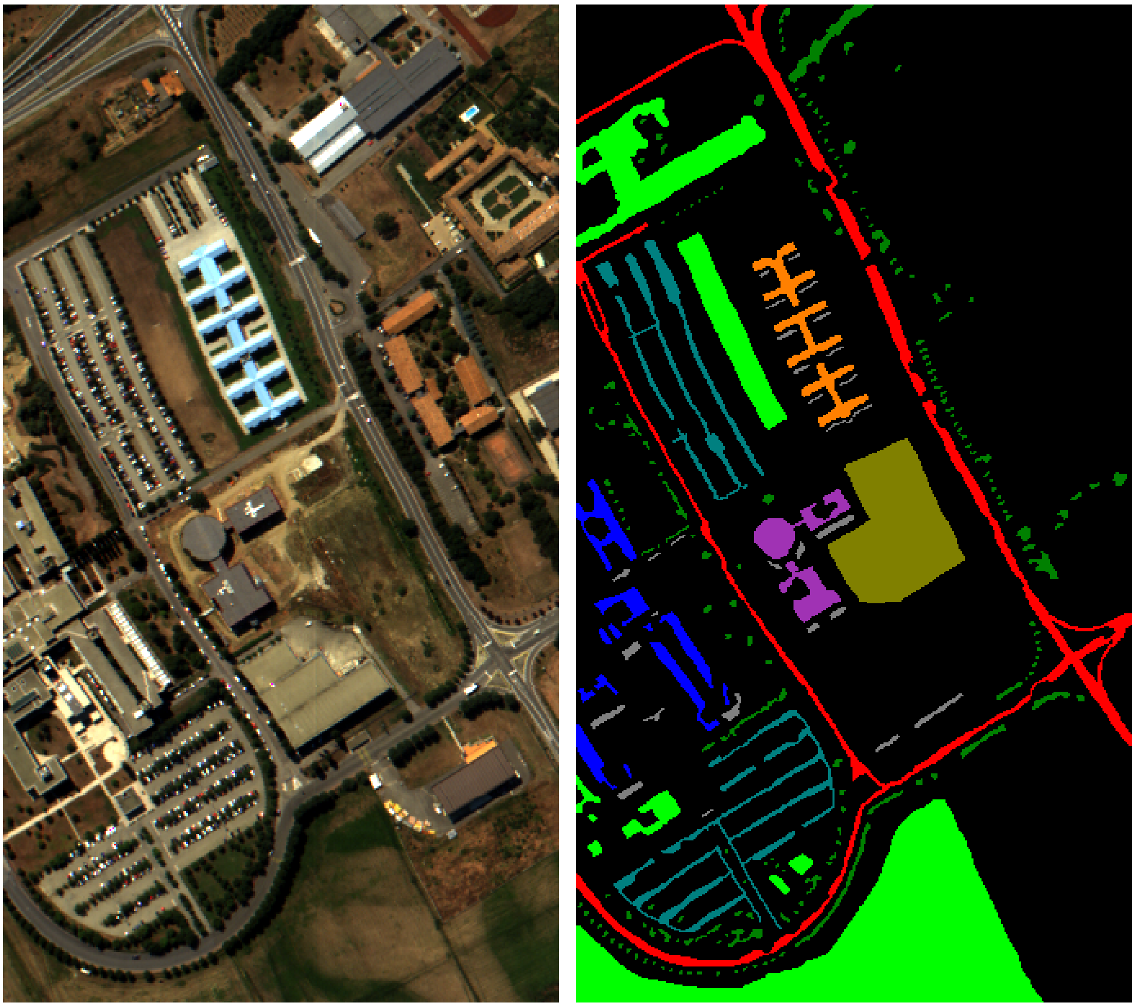
4. EXPERIMENTS

At the beginning of this chapter, the focus is on the introduction of the Pavia University dataset used in the hyperspectral image classification task. The dataset, its detailed class distribution, and statistical analysis are presented. Then, in the following sections, experiments delve into the generation of heatmaps using DeepHyperX and M3d-Cam tools, highlighting the decision-making process of the RDACN model. The section 4.3 outlines the strategies for training and partitioning the dataset into patches. The experiments are covered in three main sections to explore the impact of varying patch sizes, compare Grad-CAM to Grad-CAM++, and analyze band intensity levels through Guided Backpropagation. Subsequently, the results and discussions provide interpretations into attention map variations, XAI method comparisons, and the importance of spectral bands for different classes.

4.1 Pavia University Dataset

One of the widely used hyperspectral remote sensing image classification datasets is the Pavia University dataset. The Pavia University dataset is an image containing an urban area of size 340×610 pixels. It has nine thematic classes and has been acquired using the ROSIS-03 sensor with 1.3 m spatial resolution over the city of Pavia, Italy. The ROSIS-03 sensor has 115 data channels with a spectral coverage ranging from 0.43 to 0.86 μm. After eliminating 12 noisy bands, 103 bands have been left for processing (Fig. 4.1).

The Table 4.1 provides the detailed class distribution per class and the statistics of the Pavia University dataset, including the mean and the standard deviation. While the dataset contains a total of 42,776 labeled instances, the class 'Shadows' has the minimum number of instances, which is 947, while the class 'Meadows' has the maximum number, with 18,649 instances. The mean value of 4752.8 represents



(a) Pavia University

(b) Ground truth

Figure 4.1 Pavia University dataset, (a) color image, (b) ground truth map; classes (pixel count): 🟥 Asphalt (6631), 🟩 Trees (3064), 🟪 Bitumen (1330), 🟩 Meadows (18649), 🟧 Painted Metal sheets (1345), ▨ Shadows (947), 🟦 Gravel (2099), 🟫 Bare soil (5029) and 🟦 Self-blocking bricks (3682).

the average number of instances in the dataset, with a spread of 5540.3 instances around this mean coming from the standard deviation of the distribution.

Table 4.1 Class Distribution in the Pavia University Dataset

Class	**Instances**
Asphalt	6631
Trees	3064
Bitumen	1330
Meadows	18649
Painted Metal sheets	1345
Shadows	947
Gravel	2099
Bare soil	5029
Self-blocking bricks	3682
Total	**42776**
Mean	**4752.8**
Standard Deviation	**5540.3**
Maximum Instances	**18649**
Minimum Instances	**947**

4.2 Extracting Attention Maps

For the experiments, the Pavia University dataset partition for training, testing, and validation follows a strategy in which 95% of the ground truth samples were allocated for testing. Of the remaining 5%, its 95% is used for training purposes, leaving 107 samples for validation.

The choice behind training with a relatively modest dataset size, specifically 5%, was to prevent the risk of overfitting. This is done to avoid choosing adjacent pixels. If the training dataset were significantly larger than the dataset used for testing, there could be a problem where the testing data points are too closely aligned with the training data. Since the model has seen the instances during training it would already know which category to classify them into. This could be problematic when it is applied to unseen instances due to the generalization ability of the model. By reserving 95% of the Pavia University dataset for testing and 5% for training, the aim was to train a model with a robust evaluation and generalized capabilities. This strategy enhances its ability to capture and adapt to different patterns and trends in real-world scenarios. In order to partition the image into randomly sampled patches, we employed the DeepHyperX toolbox introduced by Audebert, Le Saux & Lefèvre (2019). This comprehensive deep learning tool, named DeepHyperX, takes the Pavia University input image and divides it into patches. Then, we fed these patches into our previously mentioned integrated RDACN model. Following the first step, we utilized the M3d-Cam library which has XAI tools provided by

Gotkowski et al. (2020) to generate attention maps based on the chosen model and XAI method. The generated attention maps and images shed light on the model's decision-making process through the selected XAI method. The generated output results are provided in the next chapter where the results are provided.

4.2.1 DeepHyperX

The deep learning toolbox for hyperspectral images, namely DeepHyperX, consists of modules written in PyTorch libraries aimed at training and comparing different benchmark deep learning models on hyperspectral images. It is composed of a variety of models, from linear support vector machines to 3D CNNs. It is implemented to train and evaluate models on various hyperspectral image datasets as well. It offers flexibility in adjusting parameters related to the size of the chosen dataset, spatial features, or optimization technique. Considering that the most recent model available in this repository is from 2018, we integrated our preferred model, the RDACN, which is more recent and demonstrates superior performance for the classification of the Pavia University dataset.

One of the key elements in training the models lies in the patching mechanism. Random patches, made of image pixels from the hyperspectral image, need to be generated to feed into the model. These patches were the input to the RDACN model. By breaking down the image into smaller and different patches, we gain insights into how different patch sizes contribute to the learning of the model. This approach was crucial because, otherwise, with a sequential patching mechanism, the model might memorize adjacent pixels composed of similar patterns rather than truly learning. Additionally, from the learning phase, we can also gain interpretations regarding the model's decision process by using the M3d-Cam repository.

4.2.2 M3d-Cam

In order to generate attention maps visualizing where the model focuses its attention, the M3d-Cam repository is employed in this thesis. Originally designed for generating attention maps for 2D and 3D medical images, the toolkit is capable of performing image classification and segmentation for both 2D and 3D data. Additionally, it offers the explainability methods, including Grad-CAM, Guided Grad-

CAM, Grad-CAM++, and the Guided Backpropagation. The resulting heatmaps referred to as attention maps, illustrate the regions in the input data that had the greatest influence on the model's prediction for a specific layer.

Since generating attention maps for 3D images is available via M3d-Cam, we applied a similar strategy for hyperspectral images. For the Guided Backpropagation, $k \times k \times 103$ images are generated per image where k is the patch size and 103 is the total number of channels. Analyzing these images allows us to evaluate channel importance and variations of focus based on the channels, detecting changes in attention across different input channels. On the other hand, for the Grad-CAM and Grad-CAM++ method, 2D images are generated where each heatmap is projected back onto the input patch with respect to a specified class. This approach makes it possible to visualize which image regions are crucial for the model to predict certain classes.

4.3 Experimental Setup

As mentioned at the beginning of this chapter, 95% of 340×610 Pavia University dataset is reserved for testing, while 95% of the remaining 5% were used for training, and the remaining samples (107 samples) were set aside for validation.

The primary task is to classify each image pixel, where each input sample is a square patch of $k \times k$ pixels centered on the pixel under consideration. Notably, no dimension reduction has been applied; hence, all 103 bands have been employed. By following this approach, a comprehensive analysis was possible due to the full spectrum of information being considered.

The experimental settings for all runs included training the model for 20 epochs. After training, we saved the weights of the model from the final epoch and utilized those weights for an additional 20 epochs, employing a batch size of 100. This decision was made to improve the model further, building upon the knowledge rooted in the initial training phase.

It is also worth emphasizing the role of the validation set. The validation set is essential for fine-tuning the remaining parameters and steering the model toward optimal model performance. Throughout the training process, with the help of the aforementioned strategies, the goal was to improve the model's accuracy and kappa scores while simultaneously gaining insights into the model's predictions.

4.4 Results & Discussion

The upcoming sections will cover three distinct experiments. The first experiment will investigate the change in attention maps for classes by varying patch sizes through the Grad-CAM++ method. Subsequently, the second experiment involves a comparative analysis of the attention maps generated by Grad-CAM and Grad-CAM++ with a fixed patch size. Finally, the last experiment will focus on band activations using Guided Backpropagation.

4.4.1 Effects of Varying Patch Sizes via Grad-CAM++

In the first experiment, we investigated the impact of varying patch sizes via Grad-CAM++, including $k = 7$, $k = 11$, $k = 15$, $k = 19$, and $k = 23$. Figure 4.2 show the differences in class visualization when applying Grad-CAM++ to each patch size.

When examining the comparison between patch size and class visualization, a general pattern emerged. Across different patch sizes, one of the key observations was the occurrence of heightened activations concentrated at the central regions of the patches. This general pattern held for the majority of classes, indicating a consistent influence of the central regions on attention.

However, there were nuances specifically within the "Asphalt" and "Self-blocking Brick" classes. For these classes, more focused and distinctive highlighted regions emerged in the attention maps, highlighting some level of differentiation. This differentiation suggests that for some classes the attention may not be uniformly concentrated on the central regions in the image for all classes; rather, there might be minor variations in the model's attention for some classes.

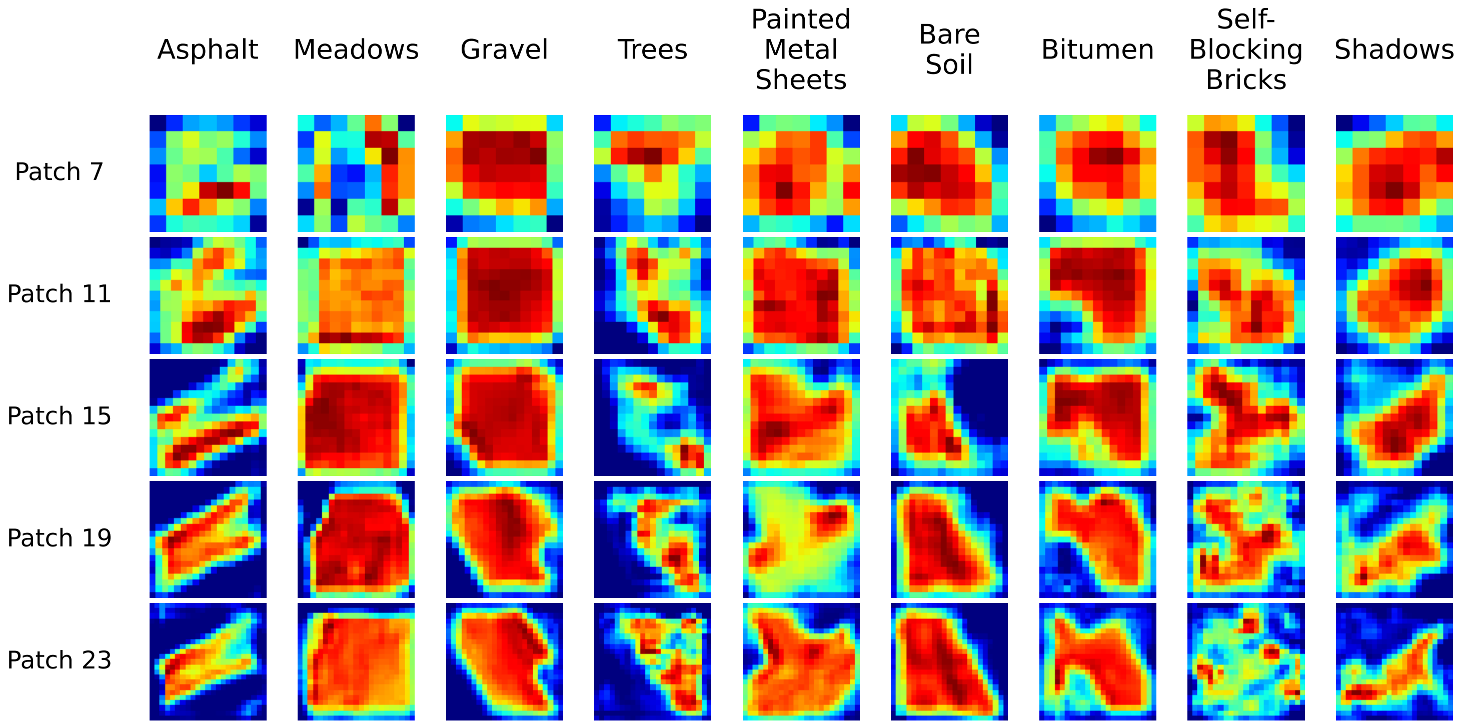


Figure 4.2 Heatmaps visualizing the spatial importance with respect to patch size using Grad-CAM++. Similar to typical Grad-CAM heatmaps, red regions represent high scores for the respective class, blue regions represent low scores, and color intensity indicates the degree of importance.

4.4.2 Grad-CAM vs Grad-CAM++

In the second experiment, in a separate comparison, we fixed patch sizes at 7, 11, 15, 19, and 23. The aim was to compare the Grad-CAM and Grad-CAM++ methods. These comparisons allowed us to assess the performance of both methods consistently across all nine classes, as shown in Figure 4.3, Figure 4.4, Figure 4.5, Figure 4.6, Figure 4.7.

While comparing the differences between the two methods, a perceptible pattern became apparent. It was observed that Grad-CAM++ exhibited a greater degree of activation and a more highlighted importance on specific regions. The visual outcomes (in Figure 4.3, Figure 4.4, Figure 4.5, Figure 4.6, Figure 4.7) reflect those differences, indicating that Grad-CAM++ is more effective in highlighting certain regions in the attention maps while operating under a fixed patch size.

Another indication of Grad-CAM++'s superiority over Grad-CAM is its ability to generate heatmaps for all classes, a capability that Grad-CAM lacked. The inability of Grad-CAM to generate heatmaps for the “Gravel” and “Shadows” is evident in Figure 4.4.

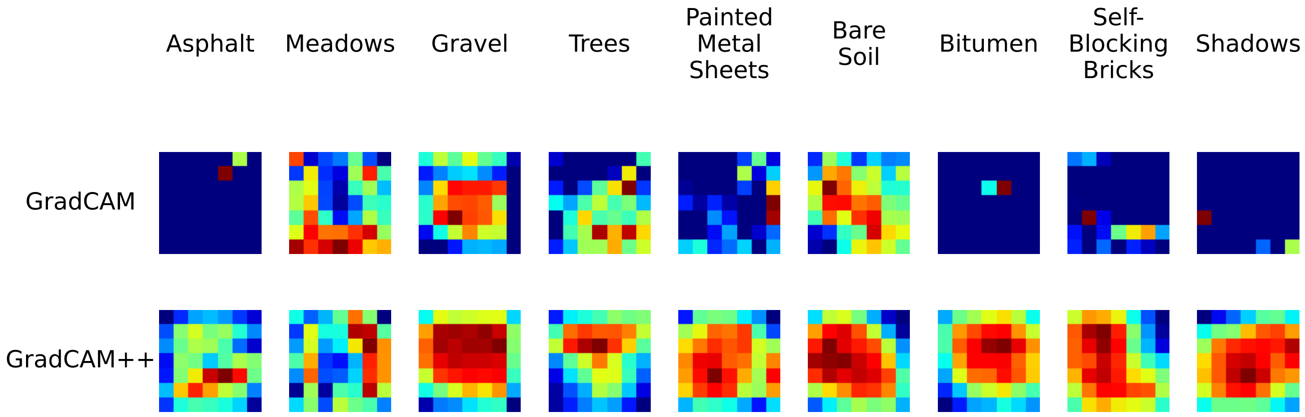


Figure 4.3 Heatmaps showing the results of Grad-CAM and Grad-CAM++ methods for a Patch Size of 7 across all nine classes in the Pavia University dataset. Red regions represent high scores for the respective class, blue regions represent low scores, and color intensity indicates the degree of importance.

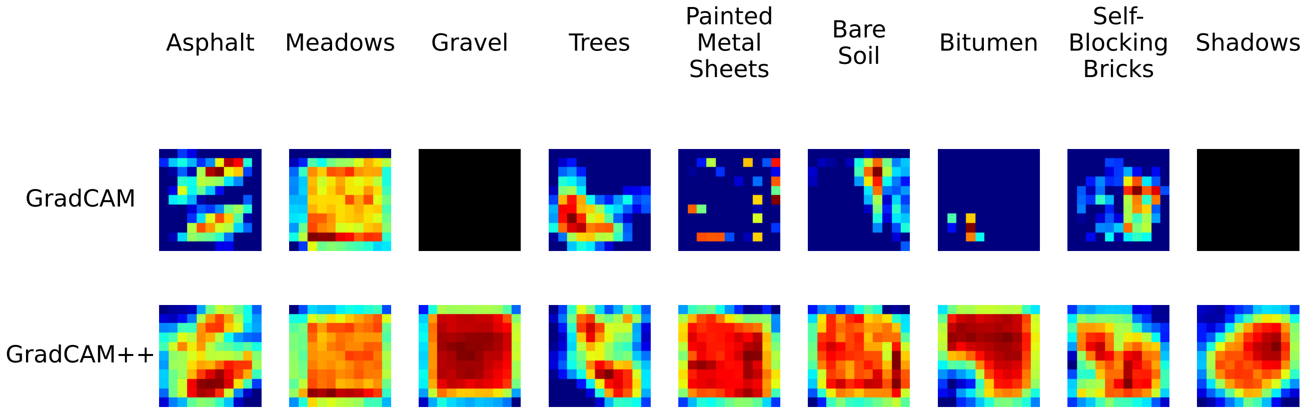


Figure 4.4 Heatmaps showing the results of Grad-CAM and Grad-CAM++ methods for a Patch Size of 11 across all nine classes in the Pavia University dataset. Red regions represent high scores for the respective class, blue regions represent low scores, and color intensity indicates the degree of importance.

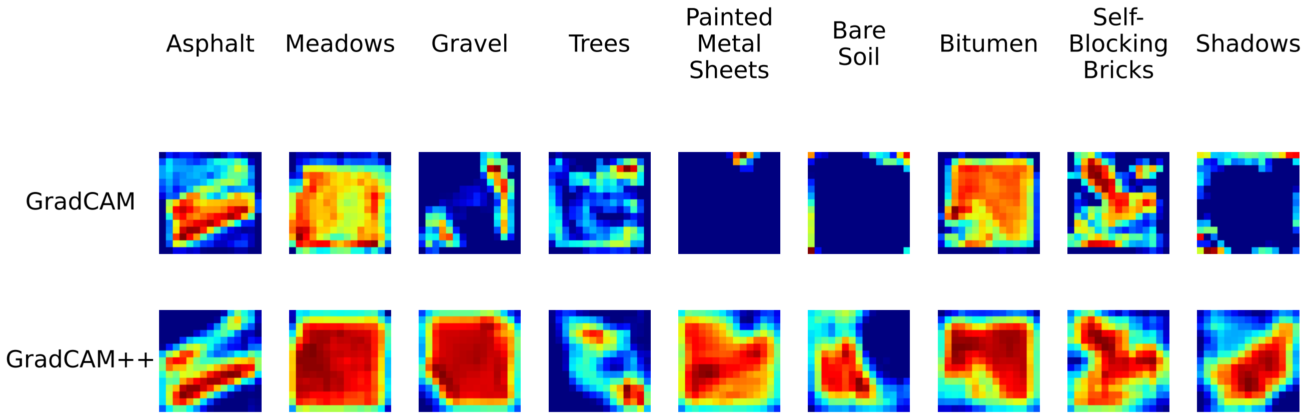


Figure 4.5 Heatmaps showing the results of Grad-CAM and Grad-CAM++ methods for a Patch Size of 15 across all nine classes in the Pavia University dataset. Red regions represent high scores for the respective class, blue regions represent low scores, and color intensity indicates the degree of importance.

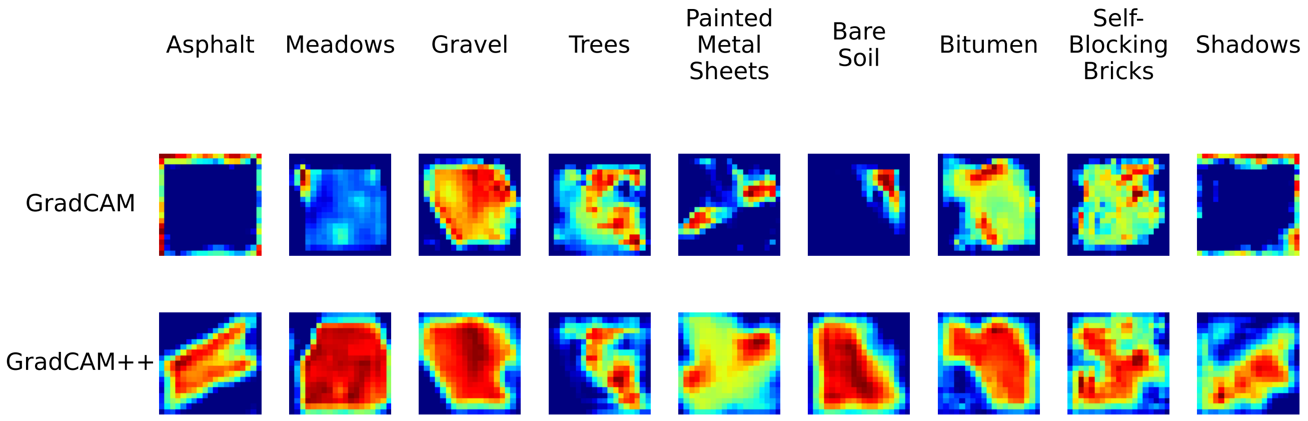


Figure 4.6 Heatmaps showing the results of Grad-CAM and Grad-CAM++ methods for a Patch Size of 19 across all nine classes in the Pavia University dataset. Red regions represent high scores for the respective class, blue regions represent low scores, and color intensity indicates the degree of importance.

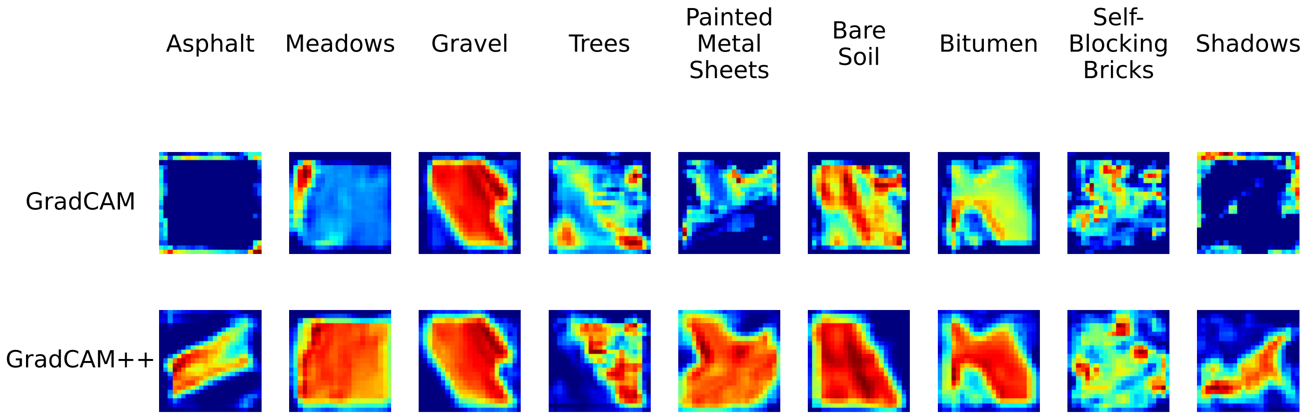


Figure 4.7 Heatmaps showing the results of Grad-CAM and Grad-CAM++ methods for a Patch Size of 23 across all nine classes in the Pavia University dataset. Red regions represent high scores for the respective class, blue regions represent low scores, and color intensity indicates the degree of importance.

4.4.3 Band Activation via Guided Backpropagation

In the third and final experiment, Guided Backpropagation was employed to calculate the average, maximum, and minimum activations for each band (Figure 4.8, Figure 4.9, Figure 4.10, Figure 4.11, Figure 4.13, Figure 4.14, Figure 4.15, Figure 4.16) of the input image patches, assessing their importance for the underlying model. The investigation of the importance of each band's activation enabled us to gain insights into the overall contribution of individual bands by averaging their activations. By looking at the average activations of individual bands, it was possible to comment on their respective band's importance.

In the error bar plots, the central points represent the mean intensity levels for the corresponding ith channel across 103 spectral bands, considering patches of size 23 for each class. Concurrently, the range surrounding these central points, referred to as error bars, conveys the variability in the data by indicating both the minimum and maximum values. Combining these three statistical components -with central points denoting the mean and the range encapsulating the intensity spread- revealed distinct trends for each class.

Observing specific classes, the "Asphalt" class exhibited spikes after the 80th spectral band (around 83 and 85) that can be seen in Figure 4.8. The "Meadows" class displayed a relatively stable trend throughout most bands, yet around the 70th and 83rd bands, spikes were observed (Figure 4.9). On the other hand, the "Gravel" class showed a declining trend around the 83rd band, followed by a spike at the 85th band (Figure 4.10). The trends for the mean intensity levels for the patches of "Trees" (Figure 4.11), "Painted Metal Sheets" (Figure 4.12), and "Self-Blocking Bricks" (Figure 4.15) classes generally remained stable, although the range between the minimum and maximum values was notably extensive.

In further observations, notable spikes after relatively stable phases examples were seen in the "Bare Soil" (Figure 4.13) and "Shadows" (Figure 4.16) classes, spikes occurring between the 70th and 83rd bands. Finally, for the "Bitumen" class, overall intensity levels were heightened between the 10th and 25th bands (Figure 4.14).

In summary, comprehensively analyzing the error bar plots generated by the Guided Backpropagation method reveals that band importance generally follows a stable pattern. However, specific bands showcased heightened activity for certain classes, providing valuable insights into the variational spectral characteristics of each class. Hence, these results enhance our understanding of how individual spectral bands contribute to the model's final prediction for each class.

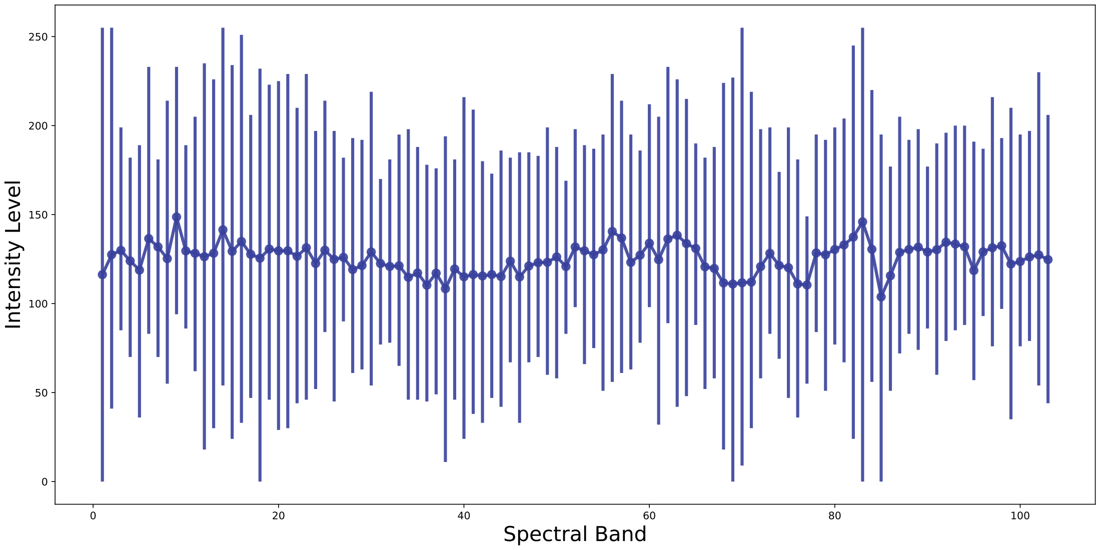


Figure 4.8 Error plot depicting Guided Backpropagation results for patches belonging to the "Asphalt" class. Error bars illustrate the range between the minimum and maximum pixel intensities across all patches in channel i. The line represents the mean intensity levels across patches of size 23.

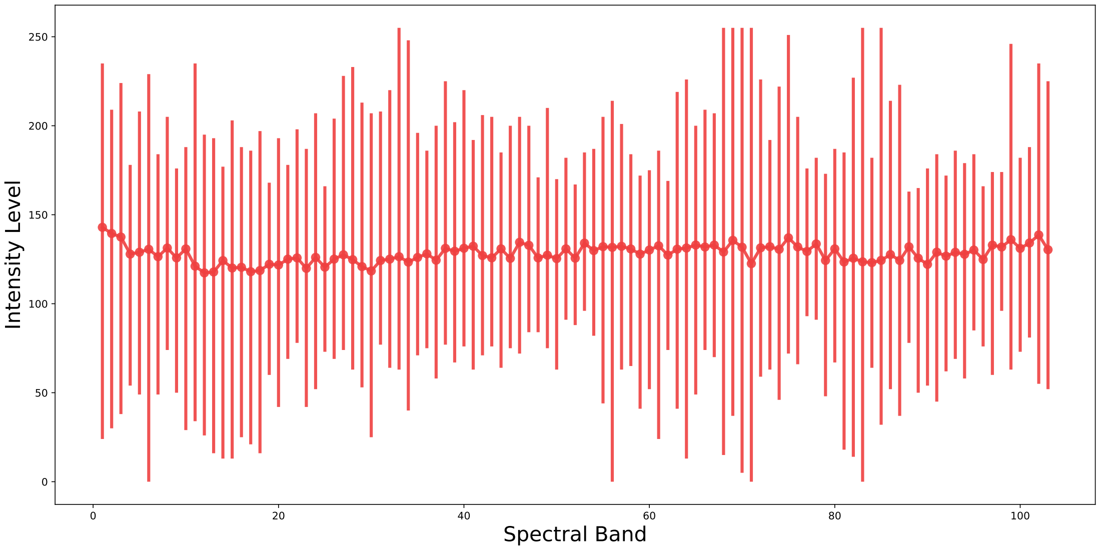


Figure 4.9 Error plot depicting Guided Backpropagation results for patches belonging to the "Meadows" class. Error bars illustrate the range between the minimum and maximum pixel intensities across all patches in channel i. The line represents the mean intensity levels across patches of size 23.

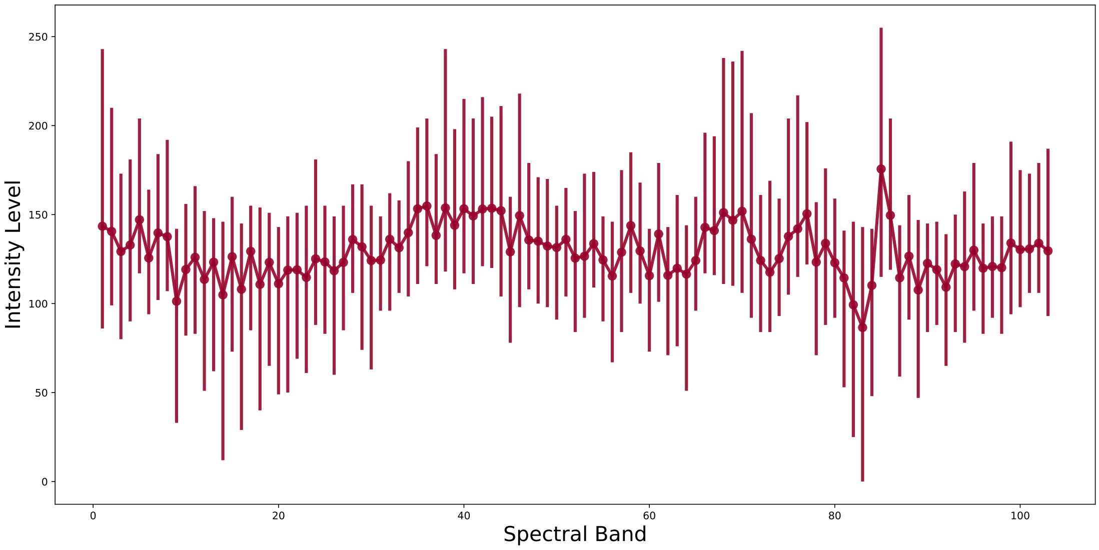


Figure 4.10 Error plot depicting Guided Backpropagation results for patches belonging to the “Gravel” class. Error bars illustrate the range between the minimum and maximum pixel intensities across all patches in channel i. The line represents the mean intensity levels across patches of size 23.

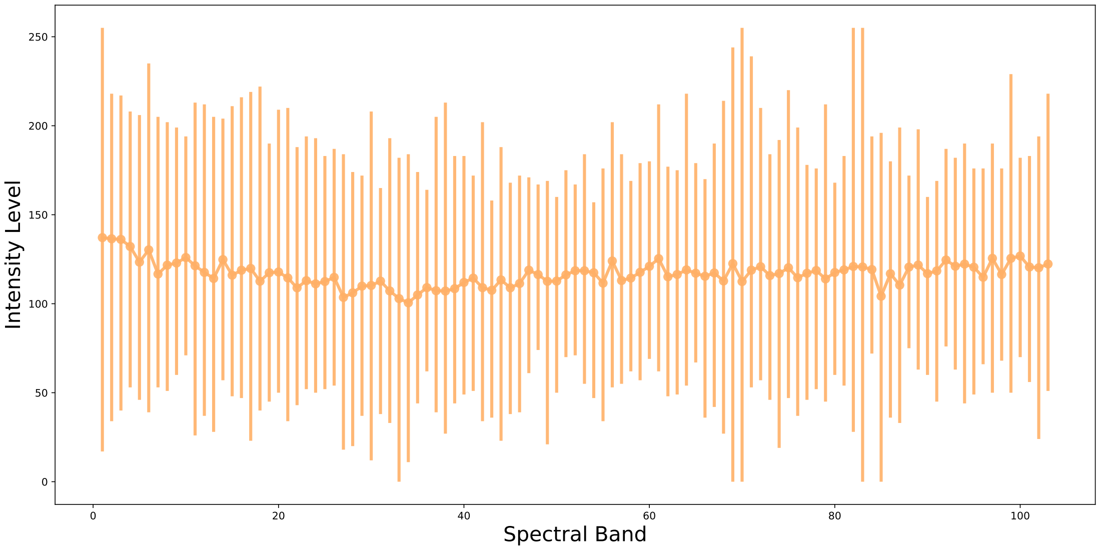


Figure 4.11 Error plot depicting Guided Backpropagation results for patches belonging to the “Trees” class. Error bars illustrate the range between the minimum and maximum pixel intensities across all patches in channel i. The line represents the mean intensity levels across patches of size 23.

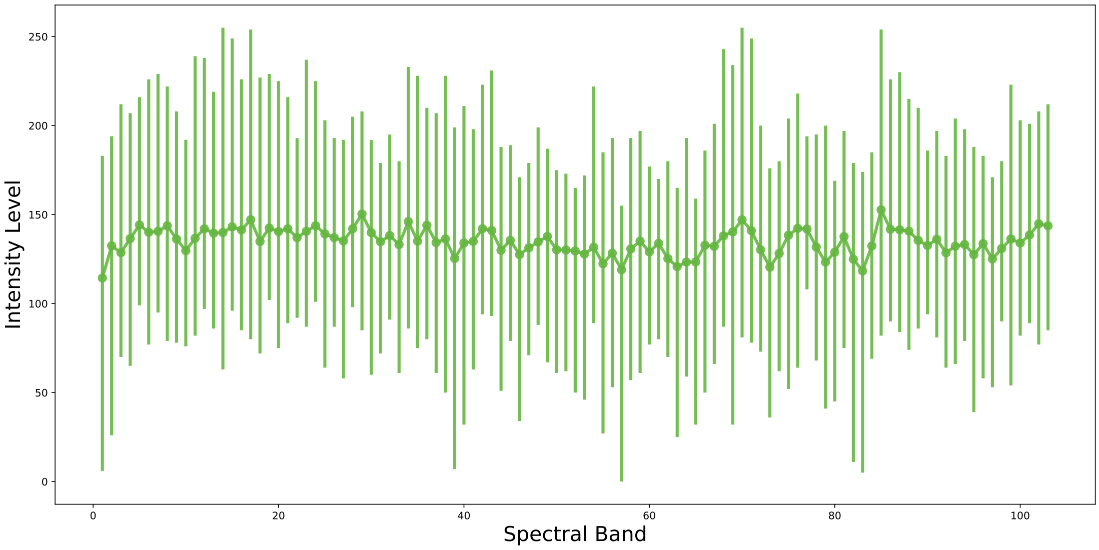


Figure 4.12 Error plot depicting Guided Backpropagation results for patches belonging to the “Painted Metal Sheets” class. Error bars illustrate the range between the minimum and maximum pixel intensities across all patches in channel i. The line represents the mean intensity levels across patches of size 23.

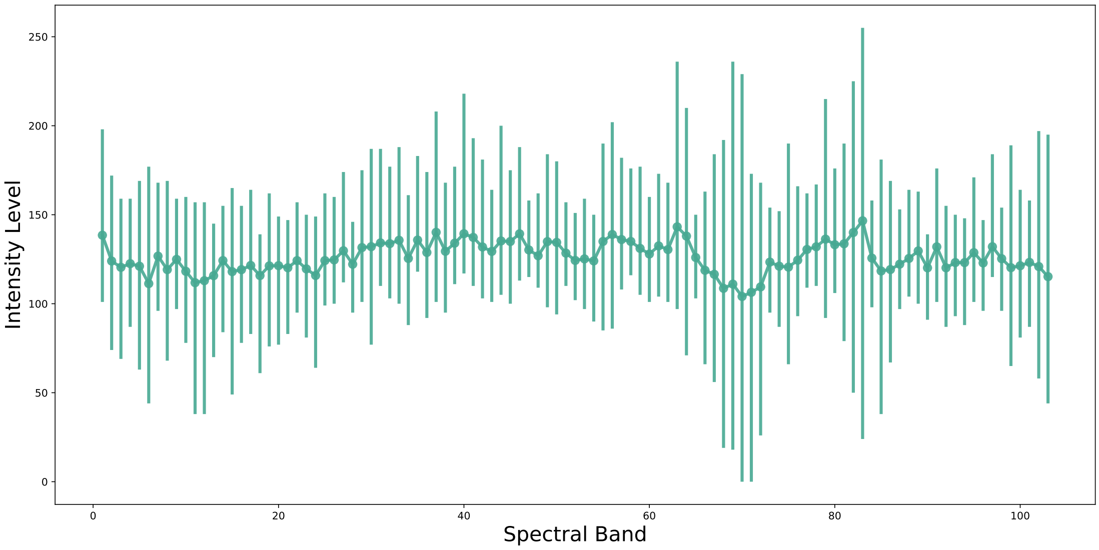


Figure 4.13 Error plot depicting Guided Backpropagation results for patches belonging to the “Bare Soil” class. Error bars illustrate the range between the minimum and maximum pixel intensities across all patches in channel i. The line represents the mean intensity levels across patches of size 23.

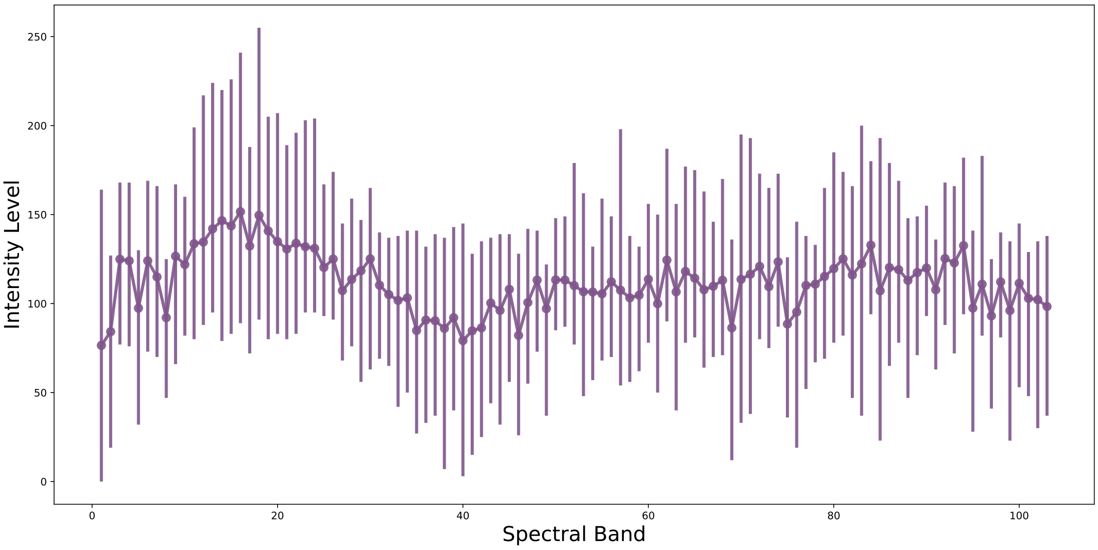


Figure 4.14 Error plot depicting Guided Backpropagation results for patches belonging to the "Bitumen" class. Error bars illustrate the range between the minimum and maximum pixel intensities across all patches in channel i. The line represents the mean intensity levels across patches of size 23.

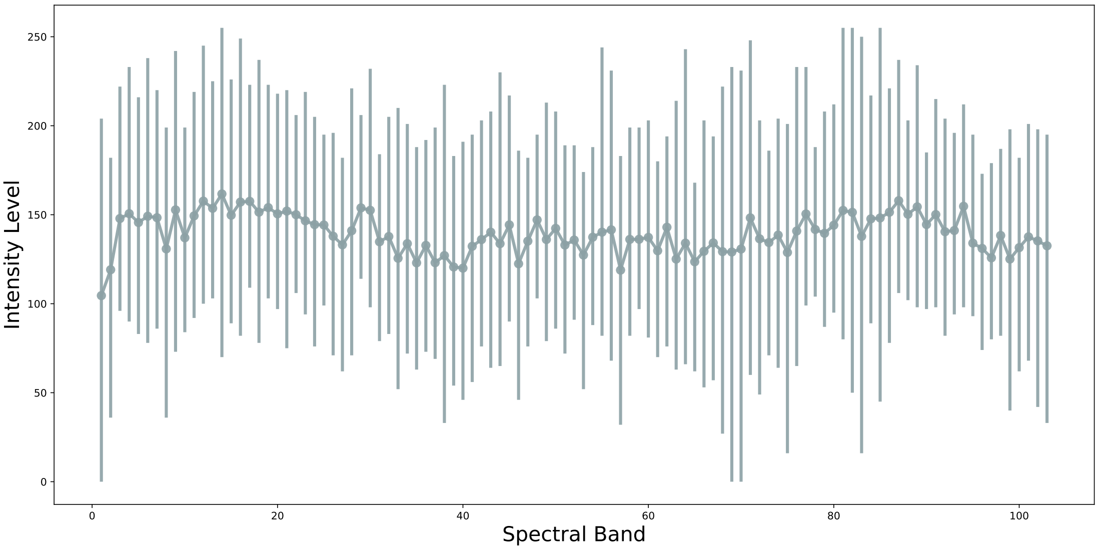


Figure 4.15 Error plot depicting Guided Backpropagation results for patches belonging to the "Self-Blocking Bricks" class. Error bars illustrate the range between the minimum and maximum pixel intensities across all patches in channel i. The line represents the mean intensity levels across patches of size 23.

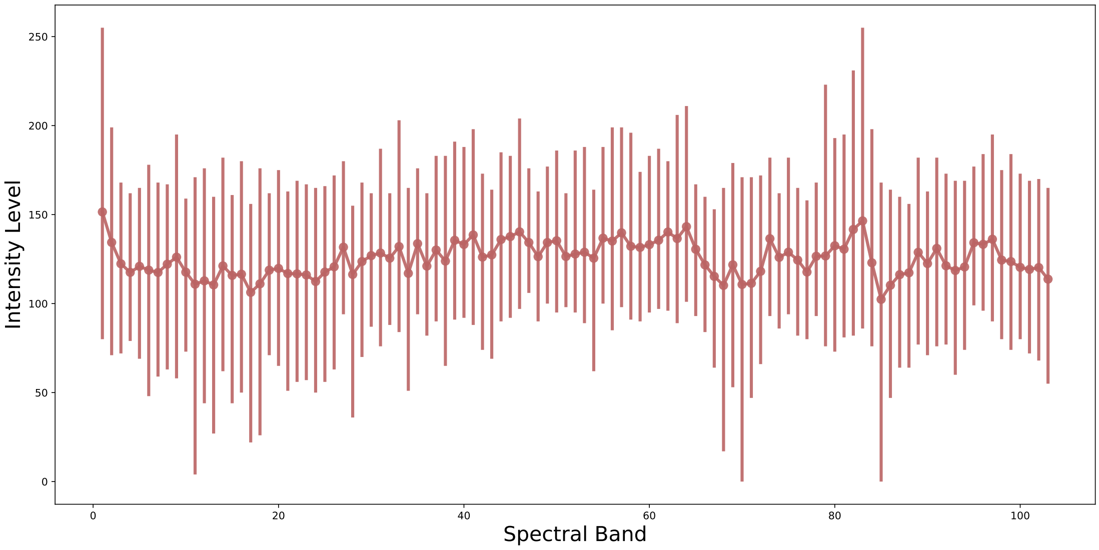


Figure 4.16 Error plot depicting Guided Backpropagation results for patches belonging to the "Shadows" class. Error bars illustrate the range between the minimum and maximum pixel intensities across all patches in channel i. The line represents the mean intensity levels across patches of size 23.

Moreover, we present the outcomes obtained through Guided Backpropagation exclusively for odd-numbered spectral bands, basing this choice on the observed similarities among the bands in the image, as illustrated in Figure 4.17. In the case of the "Asphalt" class, similar patterns were consistently emphasized across almost all bands. These highlighted patches provide crucial regions unique to their respective ground truth classes, and comprehensive results for all classes are available in Appendix B.

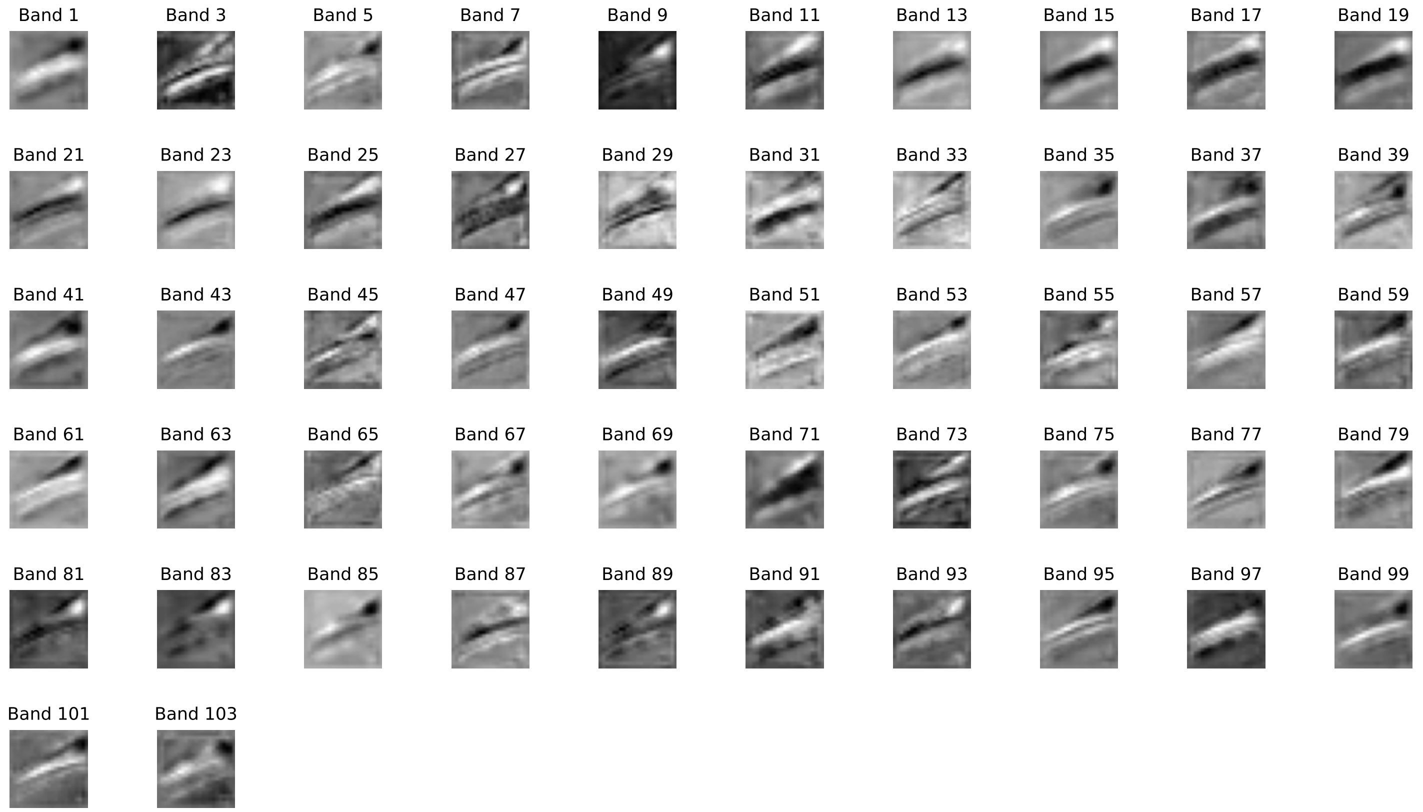


Figure 4.17 Guided Backpropagation results for “Asphalt” class across odd-numbered bands at pixel [24,83] for patch size of 23.

5. CONCLUSION

In this thesis, we investigated the inherent black-box nature of deep learning-based methodologies, which have emerged as the dominant approaches in image analysis and classification tasks. Despite their widespread use and effectiveness, these models lack transparency in their decision-making processes (Castelvecchi, 2016).

We reported results obtained through three main interpretation methods: Grad-CAM, Grad-CAM++, and Guided Backpropagation, integrated into the Residual Dense Asymmetric Convolutional Neural Network (RDACN). These explainable artificial intelligence methods provided heatmaps and images depicting the regions the RDACN model focused on during training, highlighting image patch regions that offer a category-specific interpretation of the model's decision-making.

The experiments are organized into three main sections. The first experiment explored the impact of varying patch sizes, revealing a consistent focus on central regions across nine classes and taking into account the spatial surroundings during decision-making, except for nuanced attention patterns observed in the "Asphalt" and "Self-blocking Bricks" classes. Additionally enlarging the patch size not only magnified the attention span but also resulted in distinct attention patterns, leveraging additional spatial information.

A series of experiments comparing Grad-CAM and Grad-CAM++ methods suggested that the latter is superior in terms of performance and efficiency. Grad-CAM++ exhibited heightened levels of activation and was more effective in highlighting crucial regions, even in cases where Grad-CAM failed to generate heatmaps for certain classes.

The final series of experiments, based on Guided Backpropagation, provided insights into the importance of individual spectral bands. The analysis of the results based on significance levels of bands revealed both stable patterns and specific heightened intensity spikes for certain classes, enhancing our understanding of the variational contributions of each band to the model's predictions.

In conclusion, the main contributions of this thesis are to reveal the underlying important regions and bands in image patches for hyperspectral remote sensing image classification -an under-explored field. The combination of interpretability methods and detailed experiments deepens our understanding of the model's attention dynamics, paving the way for enhanced transparency, trust, and control in deep learning-based remote sensing applications.

Future work will expand beyond the current comparative study, including new datasets and XAI methods. A particularly innovative goal will involve the expansion of XAI applications in remote sensing to the domain generalization field. This area is even less explored compared to the field of explainable hyperspectral image classification. In future research, our primary goal will be to enhance the explainability of remote sensing image classification within the context of domain generalization.

BIBLIOGRAPHY

Audebert, N., Le Saux, B., & Lefèvre, S. (2019). Deep learning for classification of hyperspectral data: A comparative review. *CoRR*, *abs/1904.10674*.

Cai, Z., Huang, Z., He, M., Li, C., Qi, H., Peng, J., Zhou, F., & Zhang, C. (2023). Identification of geographical origins of radix paeoniae alba using hyperspectral imaging with deep learning-based fusion approaches. *Food Chemistry*, *422*, 136169.

Carneiro, G. A., Pádua, L., Peres, E., Morais, R., Sousa, J. J., & Cunha, A. (2022). Segmentation as a preprocessing tool for automatic grapevine classification. In *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, (pp. 6053–6056).

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature News*, *538*(7623), 20.

Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *IEEE WACV*, (pp. 839–847).

Csurka, G., Volpi, R., & Chidlovskii, B. (2023). Semantic image segmentation: Two decades of research. 2302.06378.

Dai, W., Dai, C., Qu, S., Li, J., & Das, S. (2016). Very deep convolutional neural networks for raw waveforms. *CoRR*, *abs/1610.00087*.

Dai, Y., Jin, T., Song, Y., Sun, S., & Wu, C. (2020). Convolutional neural network with spatial-variant convolution kernel. *Remote Sensing*, *12*(17).

De Lucia, G., Lapegna, M., & Romano, D. (2022). Towards explainable ai for hyperspectral image classification in edge computing environments. *Computers and Electrical Engineering*, *103*, 108381.

Deshpande, S., Thakur, R., & Balamuralidhar, P. (2021). Learning deep spectral features for hyperspectral data using convolution over spectral signature shape. In *WHISPERS*, (pp. 1–5).

Ding, X., Guo, Y., Ding, G., & Han, J. (2019). Acnet: Strengthening the kernel skeletons for powerful CNN via asymmetric convolution blocks. *CoRR*, *abs/1908.03930*.

Gao, Y., Liu, J., Li, W., Hou, M., Li, Y., & Zhao, H. (2023). Augmented grad-cam++: Super-resolution saliency maps for visual interpretation of deep neural network. *Electronics*, *12*(23).

Gawlikowski, J., Ebel, P., Schmitt, M., & Zhu, X. X. (2022). Explaining the effects of clouds on remote sensing scene classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *15*, 9976–9986.

Gevaert, C. M. (2022). Explainable AI for earth observation: A review including societal and regulatory perspectives. *International Journal of Applied Earth Observation and Geoinformation*, *112*, 102869.

Gizzini, A. K., Shukor, M., & Ghandour, A. J. (2023). Extending cam-based xai methods for remote sensing imagery segmentation.

Gotkowski, K., González, C., Bucher, A., & Mukhopadhyay, A. (2020). M3d-cam: A pytorch library to generate 3d data attention maps for medical deep learning. *CoRR*, *abs/2007.00453*.

Gulrajani, I. & Lopez-Paz, D. (2020). In search of lost domain generalization. *CoRR*, *abs/2007.01434*.

Hacıefendioğlu, K., Demir, G., & Başağa, H. B. (2021). Landslide detection using visualization techniques for deep convolutional neural network models. *Natural Hazards: Journal of the International Society for the Prevention and Mitigation of Natural Hazards*, *109*(1), 329–350.

Huang, X., Sun, Y., Feng, S., Ye, Y., & Li, X. (2022). Better visual interpretation for remote sensing scene classification. *IEEE Geoscience and Remote Sensing Letters*, *19*, 1–5.

Iizuka, R., Xia, J., & Yokoya, N. (2024). Frequency-based optimal style mix for domain generalization in semantic segmentation of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, *62*, 1–14.

Kakogeorgiou, I. & Karantzalos, K. (2021). Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, *103*, 102520.

Kaya, G. T., Aptoula, E., & Ertürk, A. (2023). Explainable AI for Earth observation: Current methods, open challenges, and opportunities. In *Advances in Machine Learning and Image Analysis for GeoAI*. Elsevier. In preparation.

Li, S., Song, W., Fang, L., Chen, Y., Ghamisi, P., & Benediktsson, J. A. (2019). Deep learning for hyperspectral image classification: An overview. *IEEE TGRS*, *57*(9), 6690–6709.

Lundberg, S. M., Erion, G. G., & Lee, S. (2018). Consistent individualized feature attribution for tree ensembles. *CoRR*, *abs/1802.03888*.

Lundberg, S. M. & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., & Garnett, R. (Eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Meng, Z., Zhang, J., Zhao, F., Liu, H., & Chang, Z. (2022). Residual dense asymmetric convolutional neural network for hyperspectral image classification. In *IGARSS*, (pp. 3159–3162).

Nam, H., Lee, H., Park, J., Yoon, W., & Yoo, D. (2021). Reducing domain gap by reducing style bias.

Panati, C., Wagner, S., & Brüggenwirth, S. (2022). Feature relevance evaluation using grad-cam, lime and shap for deep learning sar data classification. In *2022 23rd International Radar Symposium (IRS)*, (pp. 457–462).

Ruan, Y., Dubois, Y., & Maddison, C. J. (2022). Optimal representations for covariate shift.

Sabour, S., Frosst, N., & Hinton, G. E. (2017). Dynamic routing between capsules. *CoRR*, *abs/1710.09829*.

Sahin, I., Erturk, A., & Aptoula, E. (2023). Band-based interpretability with shap for hyperspectral classification.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, (pp. 618–626).

Shapley, L. S. (1953). *17. A Value for n-Person Games*, (pp. 307–318). Princeton: Princeton University Press.

Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2015). Striving for simplicity: The all convolutional net. 1412.6806.

Su, Q., Zhang, X., Xiao, P., Li, Z., & Wang, W. (2022). Which cam is better for extracting geographic objects? a perspective from principles and experiments. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *15*, 5623–5635.

Su, S., Cui, Z., Guo, W., Zhang, Z., & Yu, W. (2022). Explainable analysis of deep learning methods for sar image classification. In *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, (pp. 2570–2573).

Sumbul, G., Charfuelan, M., Demir, B., & Markl, V. (2019). Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, (pp. 5901–5904).

Sumbul, G., de Wall, A., Kreuziger, T., Marcelino, F., Costa, H., Benevides, P., Caetano, M., Demir, B., & Markl, V. (2021). Bigearthnet-mm: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, *9*(3), 174–180.

Sun, B. & Saenko, K. (2016). Deep CORAL: correlation alignment for deep domain adaptation. *CoRR*, *abs/1607.01719*.

Taskin, G. (2022). A model distillation approach for explaining black-box models for hyperspectral image classification. In *IGARSS*, (pp. 3592–3595).

Tong, W., Chen, W., Han, W., Li, X., & Wang, L. (2020). 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.

Tuia, D., Persello, C., & Bruzzone, L. (2016). Domain adaptation for the classification of remote sensing data: An overview of recent advances. *IEEE Geoscience and Remote Sensing Magazine*, *4*(2), 41–57.

Turan, D. E., Aptoula, E., Ertürk, A., & Taskin, G. (2023). Interpreting hyperspectral remote sensing image classification methods via explainable artificial intelligence. In *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, (pp. 5950–5953).

Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., & Hu, X. (2020). Score-cam: Score-weighted visual explanations for convolutional neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, (pp. 111–119).

Wang, Y., Abliz, A., Ma, H., Liu, L., Kurban, A., Halik, U., Pietikäinen, M., & Wang, W. (2022). Hyperspectral estimation of soil copper concentration based on improved tabnet model in the eastern junggar coalfield. *IEEE TGRS*, *60*, 1–20.

Xu, Y., Sun, H., Chen, J., Lei, L., Ji, K., & Kuang, G. (2021). Adversarial self-supervised learning for robust sar target recognition. *Remote Sensing*, *13*(20).

Zhang, J., Zhao, H., & Li, J. (2021). Trs: Transformers for remote sensing scene classification. *Remote Sensing*, *13*(20).

Zhao, J., Zhang, Z., Yao, W., Datcu, M., Xiong, H., & Yu, W. (2020). Opensarurban: A sentinel-1 sar image dataset for urban interpretation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 187–203.

Zhao, L., Zeng, Y., Liu, P., & He, G. (2020). Band selection via explanations from convolutional neural networks. *IEEE Access*, *8*, 56000–56014.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2015). Learning deep features for discriminative localization.

Zhou, K., Liu, Z., Qiao, Y., Xiang, T., & Loy, C. C. (2023). Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(4), 4396–4415.

Zhu, S., Wu, C., Du, B., & Zhang, L. (2023). Style and content separation network for remote sensing image cross-scene generalization. *ISPRS Journal of Photogrammetry and Remote Sensing*, *201*, 1–11.

APPENDIX A

The motivation and preliminary results of an additional study relevant to XAI in remote sensing are presented here. It concerns the extension and adaptation of explainable artificial intelligence into domain generalization techniques within remote sensing, particularly for the Big Earth dataset (Sumbul, de Wall, Kreuziger, Marcelino, Costa, Benevides, Caetano, Demir & Markl, 2021). Although the explainable adaptation is not completed, where the goal was to understand the influences of source domains for final model prediction, the first exploratory steps in domain generalization in remote sensing took place in this thesis. Hence, we outline our plans and present the results of image classification scores in remote sensing across various domains. Finally, we conclude with insights gained from this thesis and discuss future work in this emerging field.

A.1 Domain Generalization in Remote Sensing

Domain generalization in remote sensing is a relatively under-explored yet an area that is open to progress which addresses the challenge of adapting models trained on source geographic domains to perform well in diverse and unseen target domains. The problem is challenging because remote sensing images cover large amounts of data in a single pixel, and they tend to vary significantly across different locations due to variations in terrain, climate, sensor type, and land cover types (Tuia, Persello & Bruzzone, 2016). Consequently, statistical learning methods often struggle to generalize effectively to unseen target domains for which they were not initially trained. This occurs because they assume that the source and target data have independent and identically distributed patterns, excluding the general case of out-of-distribution (Zhou, Liu, Qiao, Xiang & Loy, 2023). The aim of domain generalization is to enhance the robustness and adaptability of models by developing techniques to transfer knowledge from multiple domains to another. In this thesis, we focused on the primary domain generalization methods that are used in this work, and in the following chapters, we present their results on a remote sensing image dataset.

At the beginning of this thesis, our aim was to develop an explainable domain generalization framework capable of identifying the contribution of each domain to the final model prediction for remote sensing image classification. For this purpose, we planned a clustering mechanism involving a deep neural network with the objective of classifying input images into their respective classes based on specific domains. Following this, the network would be trained with data from all domains to generate domain-specific features. These domain-specific features would then be clustered to form domain-specific representations. When presenting the network with an unknown domain composed of similar classes, the distance between the new domain features and those of the source domain features would be calculated with a distance measure. This distance would indicate which information from the source domains contributes to the final prediction in the target domain. A smaller distance would imply a more significant involvement of features from that source domain in the final prediction. Due to time limitations, the explainable clustering part could not be completed.

Understanding the contribution of domains to the final model prediction is crucial for detecting potential biases toward specific domains. Identifying such biases could help developers in mitigating them, and ensuring fair predictions and well-generalized models. These interpretations might also prioritize the need for further data collection and refinement. However, initial steps in domain generalization for remote sensing have been taken in this thesis, as explained in the following sections.

One of the methods used in our proposed domain generalization network refers to the paper titled “Correlation Alignment for Deep Domain Adaptation” (Sun & Saenko, 2016). The authors presented an extended version of the CORAL unsupervised domain adaptation method. The original CORAL aligns the second-order statistics of the source and target distributions, while Deep CORAL can be easily integrated into different layers in CNNs or network architectures and optimized efficiently. The authors introduced a new loss function called CORAL loss to minimize the difference in learned feature covariances across domains. In contrast, the original CORAL method mitigates domain shift by aligning the second-order statistics of the source and target distributions. By minimizing the CORAL loss, Deep CORAL ensures that the learned features are both discriminative and minimize the distance between the source and target domains.

A.1.1 Optimal Representations for Covariate Shift (CAD & CondCAD)

In the work of Ruan, Dubois & Maddison (2022), within the context of domain generalization, they introduced a methodology for learning robust representations that can generalize well to unseen target domains even when there are distribution shifts between the source and target domains. Additionally, they proposed self-supervised learning objectives that leverage unlabeled data and augmentations to train robust representations achieving robustness on CLIP and state-of-the-art performance on DomainBed.

The domain generalization process typically includes two stages: learning an encoder and learning a predictor. During the first phase, the encoder is trained to map input data to representations, with the aim of learning representations that are robust to domain shifts. The second phase has a predictor that is trained to map the learned representations to the target labels using standard risk minimization techniques (Ruan et al., 2022). The authors introduced the concept of idealized domain generalization (IDG) risk to ensure the designed robust representation allows predictors trained on the source domain to perform well on the target domain (Ruan et al., 2022). The IDG risk measures the expected worst-case target risk over all possible source risk minimizers. The representation that minimizes the IDG risk is considered as optimal.

In IDG, the learner is assumed to have access to the source population risk. According to theoretical results, optimal domain generalization requires information about the target domain or representations might not uniformly outperform a constant representation without such information. Overall, the authors proposed a method for learning robust representations in domain generalization, even in the presence of distribution shifts between domains. They provided both theoretical and practical approaches focused on self-supervised learning for better generalization to unseen target domains.

Regarding the domain shift problem, in which models struggle to generalize well to unseen target domains, Zhu, Wu, Du & Zhang (2023) introduced cross-scene generalization for remote sensing by proposing the Style and Content Separation Network (SCSN). To enhance generalization, it utilizes style normalization, and the easily adaptable Style and Content Separation (SCS) module focuses on content information. Furthermore, the authors provided a separation loss that fine-tunes the network's behavior for superior performance in cross-scene generalization tasks. The domain shift problem is also addressed in the semantic segmentation task in the study by Iizuka, Xia & Yokoya (2024). The authors proposed the frequency-based optimal style mix (FOSMix) for the land cover semantic segmentation task. FOSMix is composed of three stages: a full mix in the frequency domain for maximizing style mixing, an optimal mix that introduces selective randomness for frequencies unnecessary for segmentation, and consistency regularization that guarantees stable learning for the model across various images sharing similar semantics.

A.1.2 Style Agnostic Networks (SagNet)

As frequently encountered in domain generalization challenges, SagNet was proposed to address domain shift in CNNs, which refers to a decrease in performance when faced with new test domains. The authors introduced SagNet to mitigate the style bias in CNNs by separating style encodings from class categories and focusing more on contents. The architecture involves training separate content-biased and style-biased networks on top of a feature extractor. Through the introduction of style randomization in a latent space, the content-biased network mainly focuses on content. On the other hand, the style-biased network is directed to focus on styles in an opposite manner, by adversarially disentangling them from class categories. During testing, predictions are generated by combining the feature extractor with the content-biased network, resulting in a significant reduction in style bias. Notably, without the need for domain labels or multiple domains, SagNet only controls the intrinsic bias of CNNs, which is not only adaptable to practical situations where domain boundaries are uncertain or unclear but also has the potential to enhance existing methods (Nam, Lee, Park, Yoon & Yoo, 2021). Overall, SagNet is robust against domain shift arising from variations in style across different domains and relies more on content than style.

A.2 Domain Generalization Comparisons for Remote Sensing Image

Classification

The BigEarthNet dataset (Sumbul, Charfuelan, Demir & Markl, 2019), (Sumbul et al., 2021) is a large-scale remote sensing dataset designed for land cover classification. It includes 590.326 pairs of Sentinel-1 Synthetic Aperture Radar (SAR) and Sentinel-2 multispectral image patches between June 2017 and May 2018 from 10 countries: Austria, Belgium, Finland, Ireland, Kosovo, Lithuania, Luxembourg, Portugal, Serbia, and Switzerland. The atmospheric correction for all tiles was performed using the Sentinel-2 Level 2A product generation and formatting tool.

After obtaining the medium-sized training BigEarthNet dataset, which comprises 25,000 images, the B02, B03, and B04 channels are extracted from the Sentinel-2 image patches, resulting in the finalized version named BigEarthRGB. These channels are then used as inputs for various models within the DomainBed framework that includes benchmark datasets and algorithms designed for domain generalization (Gulrajani & Lopez-Paz, 2020). Subsequently, the labels are converted to a 19-label convention based on the study of Sumbul et al. (2021). Country names are then extracted based on their geographical coordinates, and band normalization is adapted from official works, including the study of Sumbul et al. (2021). These countries are then employed as image domains for a domain generalization task centered around country distinctions. To the best of our knowledge, this represents the first attempt to classify scenes based on countries within the field of remote sensing domain generalization.

DerenNet: A Domain Generalization Framework

DerenNet is a framework designed for domain generalization in remote sensing image classification. This framework manages data from 10 different countries with diverse distributions. DerenNet utilizes a ResNet-based featurizer to extract domain-invariant features (Figure A.1). The classifier takes as input the features extracted by the featurizer. The featurizer architecture adapts the ResNet-50 model, adjusting input size and channels. Notably, DerenNet employs multitask learning, concurrently optimizing image label classification and domain classification, with a custom classifier replacing the fully connected layer for feature extraction.

The ResNet-50 architecture is used to retrieve label and domain features from input batches. Then, the covariance matrices for label features are used to calculate the CORAL loss between feature matrices of batches. Ideally, to minimize the CORAL loss, the batch size should be higher to contain images from all classes for all domains. Additionally, it calculates domain-specific features later for use in XAI part in future studies.

During training, it uses cross-entropy loss for label classification on every image, domain classification loss, and a covariance alignment penalty for encouraging domain-invariant feature learning. Since the weights are not shared between the domain and label featurizers, their losses are not affected by each other. As mentioned earlier, the classification loss belonging to domain classifiers is designed to be used for future XAI purposes to retrieve domain-specific features. DerenNet's motivation stems from its ability to balance label classification accuracy with domain-invariant feature learning while simultaneously retrieving domain-specific features.

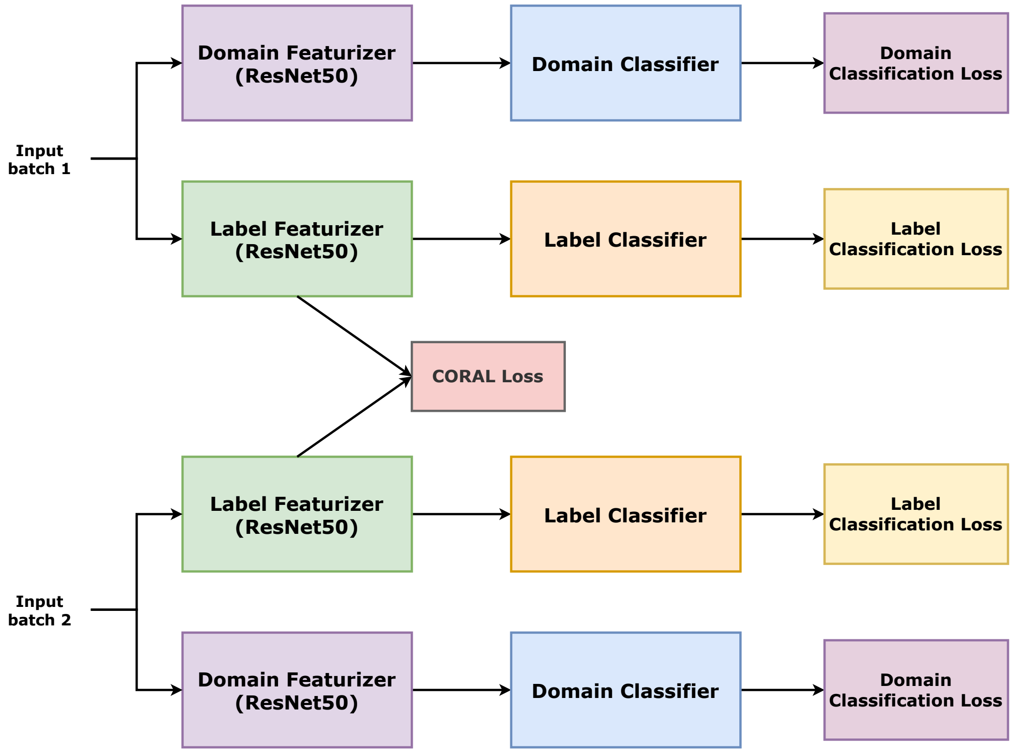

Domain Featurizer (ResNet50)
Domain Classifier
Domain Classification Loss
Input batch 1
Label Featurizer (ResNet50)
Label Classifier
Label Classification Loss
CORAL Loss
Label Featurizer (ResNet50)
Label Classifier
Label Classification Loss
Input batch 2
Domain Featurizer (ResNet50)
Domain Classifier
Domain Classification Loss

Figure A.1 DerenNet: A domain generalization framework (example illustration for two input batches)

The oracle selection method, utilizing the accuracy from the test-domain validation set aligned with the test domain's distribution, yields the optimal model performance. For hyperparameter tuning during evaluations, a random search is performed for each model using a split of 80% for training and 20% for testing on the data from each domain. The columns associated with country names in Table A.1 and Table A.2 represent the target domains during testing, and each row value within these columns displays the average of the classification accuracy scores along with its respective standard error for one of the four algorithms. Additionally, the averages of accuracy scores of all countries as target domains per algorithm are given separately in Table A.4 for the oracle selection method. As anticipated, given its access to the test domain during testing, this method establishes the upper bound for the experiments. CondCAD stands out as the leading model, achieving an average accuracy of 23.0. On the other hand, the training-domain validation set follows a different approach in that each training domain is partitioned into training and validation subsets. Then, validation subsets of each training domain are pooled to construct the final validation set. The averages of accuracy scores of all countries as target domains per algorithm are also given for this testing method separately

Table A.1 BigEarthRGB Model Evaluation for Training Domain Validation Set model selection

	Dataset: BigEarthRGB, model selection method: training-domain validation set										
Algorithm	Austria	Belgium	Finland	Ireland	Kosovo	Lithuania	Luxembourg	Portugal	Serbia	Switzerland	Avg
SagNet	**14.3 +/- 0.8**	**10.4 +/- 0.3**	20.9 +/- 1.4	**35.6** +/- 1.0	1.4 +/- 0.8	34.5 +/- 3.0	**30.6 +/- 0.8**	10.3 +/- 1.4	6.9 +/- 0.6	0.9 +/- 0.3	16.6
CAD	12.9 +/- 1.3	9.7 +/- 0.2	17.4 +/- 1.9	29.9 +/- 0.9	0.8 +/- 0.5	37.0 +/- 1.9	29.1 +/- 2.0	**11.4 +/- 1.8**	8.4 +/- 1.3	**1.5 +/- 0.9**	15.8
CondCAD	13.1 +/- 0.5	9.5 +/- 0.5	21.4 +/- 0.7	31.5 +/- 1.1	1.1 +/- 0.7	34.3 +/- 2.9	27.0 +/- 2.3	8.8 +/- 0.9	**9.7 +/- 0.6**	0.7 +/- 0.5	15.7
DerenNet	12.8 +/- 1.3	9.8 +/- 0.7	**24.8 +/- 2.8**	31.5 +/- 1.3	**1.9 +/- 0.9**	**38.3** +/- 2.2	30.0 +/- 2.1	10.6 +/- 1.2	8.5 +/- 1.3	0.8 +/- 0.5	**16.9**

Table A.2 BigEarthRGB Model Evaluation for Test Domain Validation Set (Oracle) model selection

	Dataset: BigEarthRGB, model selection method: test-domain validation set (oracle)										
Algorithm	Austria	Belgium	Finland	Ireland	Kosovo	Lithuania	Luxembourg	Portugal	Serbia	Switzerland	Avg
SagNet	16.8 +/- 0.3	9.8 +/- 0.1	29.8 +/- 1.1	35.0 +/- 0.9	3.6 +/- 0.8	38.5 +/- 0.8	**32.9 +/- 1.3**	13.4 +/- 0.6	12.7 +/- 2.0	3.8 +/- 0.8	19.6
CAD	28.0 +/- 3.7	10.9 +/- 0.6	**31.4 +/- 0.9**	**43.5 +/- 1.6**	**5.0 +/- 2.1**	46.8 +/- 2.2	31.7 +/- 1.7	**14.7 +/- 1.1**	11.1 +/- 0.6	3.6 +/- 1.0	22.7
CondCAD	**32.7 +/- 1.6**	**11.7 +/- 0.3**	27.4 +/- 0.5	38.7 +/- 2.2	4.2 +/- 1.3	**50.4 +/- 2.0**	31.0 +/- 1.1	14.3 +/- 0.6	**14.3 +/- 1.3**	**5.0 +/- 0.7**	**23.0**
DerenNet	16.0 +/- 0.8	10.1 +/- 0.5	28.5 +/- 1.7	30.7 +/- 0.9	4.2 +/- 1.0	38.2 +/- 3.1	31.3 +/- 1.3	13.9 +/- 0.2	10.0 +/- 0.6	2.4 +/- 0.8	18.5

Table A.3 Test Averages for Training Domain Validation Set model selection method

Averages, model selection method: training-domain validation set		
Algorithm	BigEarthRGB	Avg
SagNet	16.6 +/- 0.5	16.6
CAD	15.8 +/- 0.4	15.8
CondCAD	15.7 +/- 0.4	15.7
DerenNet	**16.9 +/- 0.3**	**16.9**

Table A.4 Test Averages for Test-Domain Validation (Oracle) model selection method

Averages, model selection method: test-domain validation (oracle)		
Algorithm	BigEarthRGB	Avg
SagNet	19.6 +/- 0.4	19.6
CAD	22.7 +/- 0.9	22.7
CondCAD	**23.0 +/- 0.4**	**23.0**
DerenNet	18.5 +/- 0.2	18.5

in Table A.3. It is noteworthy that, with this testing method, DerenNet emerges as the top-performing model on average.

Currently, our initial attempts in this direction include the application of a remote sensing dataset for use in a domain generalization task, indicating the early steps of exploration. In future work, we aim to enhance our proposed DerenNet model and integrate XAI to improve the interpretation of the influence of source domains on final predictions while improving the accuracy scores for target domains in the task of remote sensing image classification.

APPENDIX B

Guided Backpropagation results for target classes across odd-numbered channels for a patch size of 23

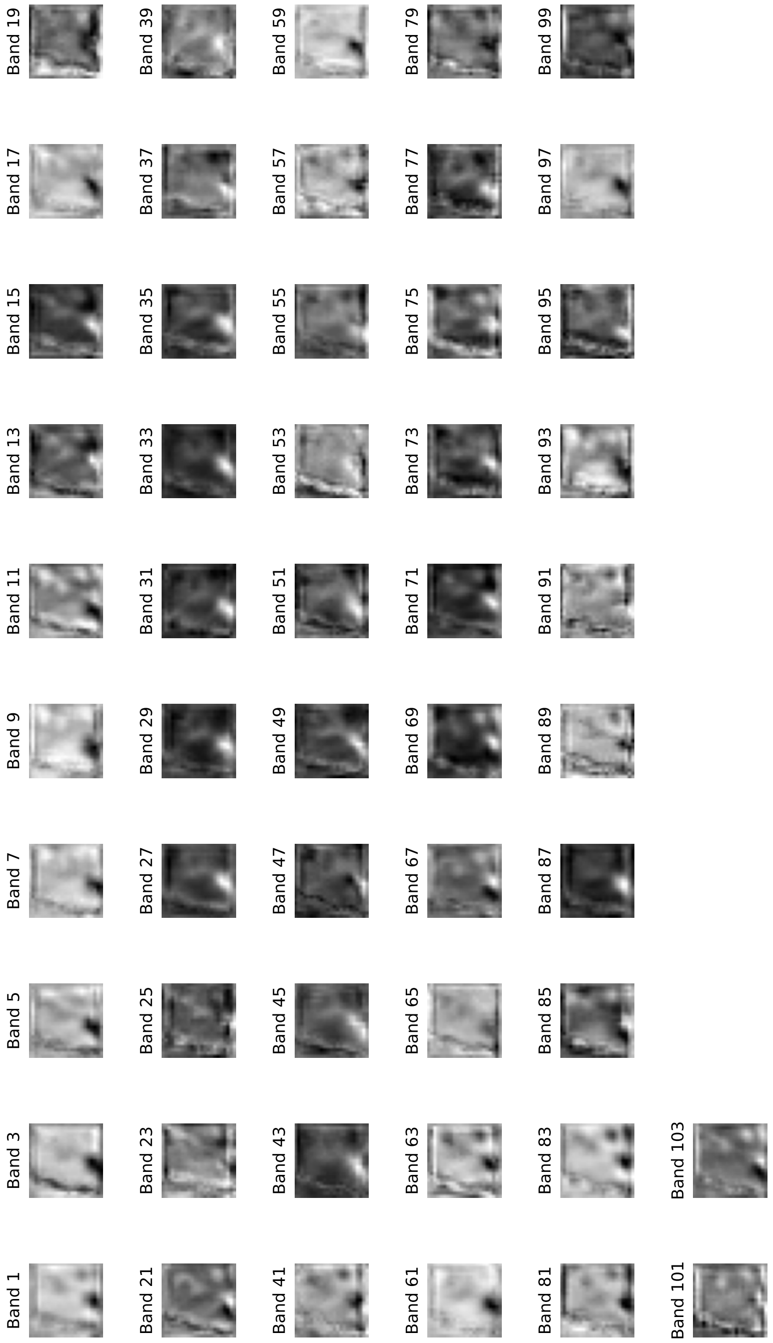


Figure B.1 Guided Backpropagation results for “Meadows” class across odd-numbered bands.

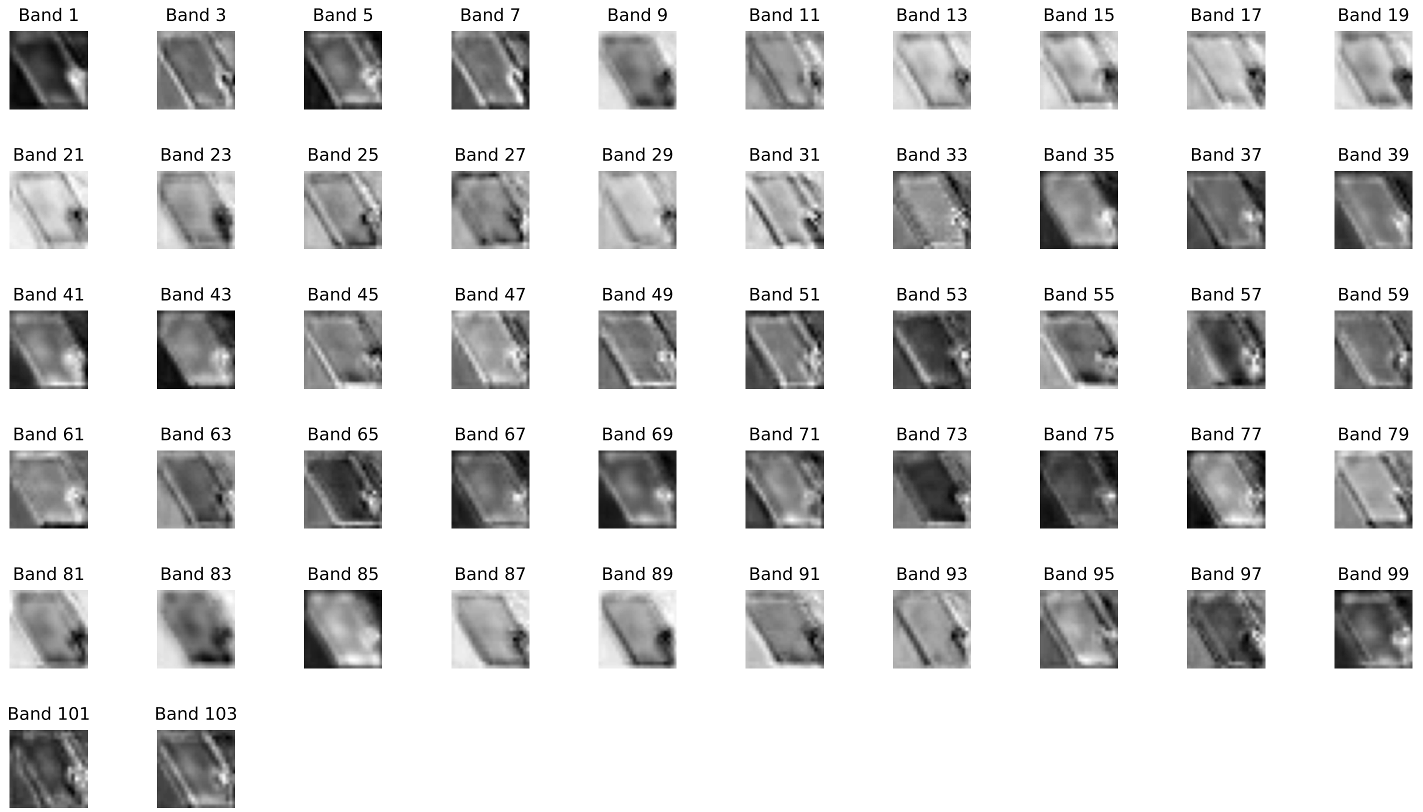


Figure B.2 Guided Backpropagation results for “Gravel” class across odd-numbered bands.

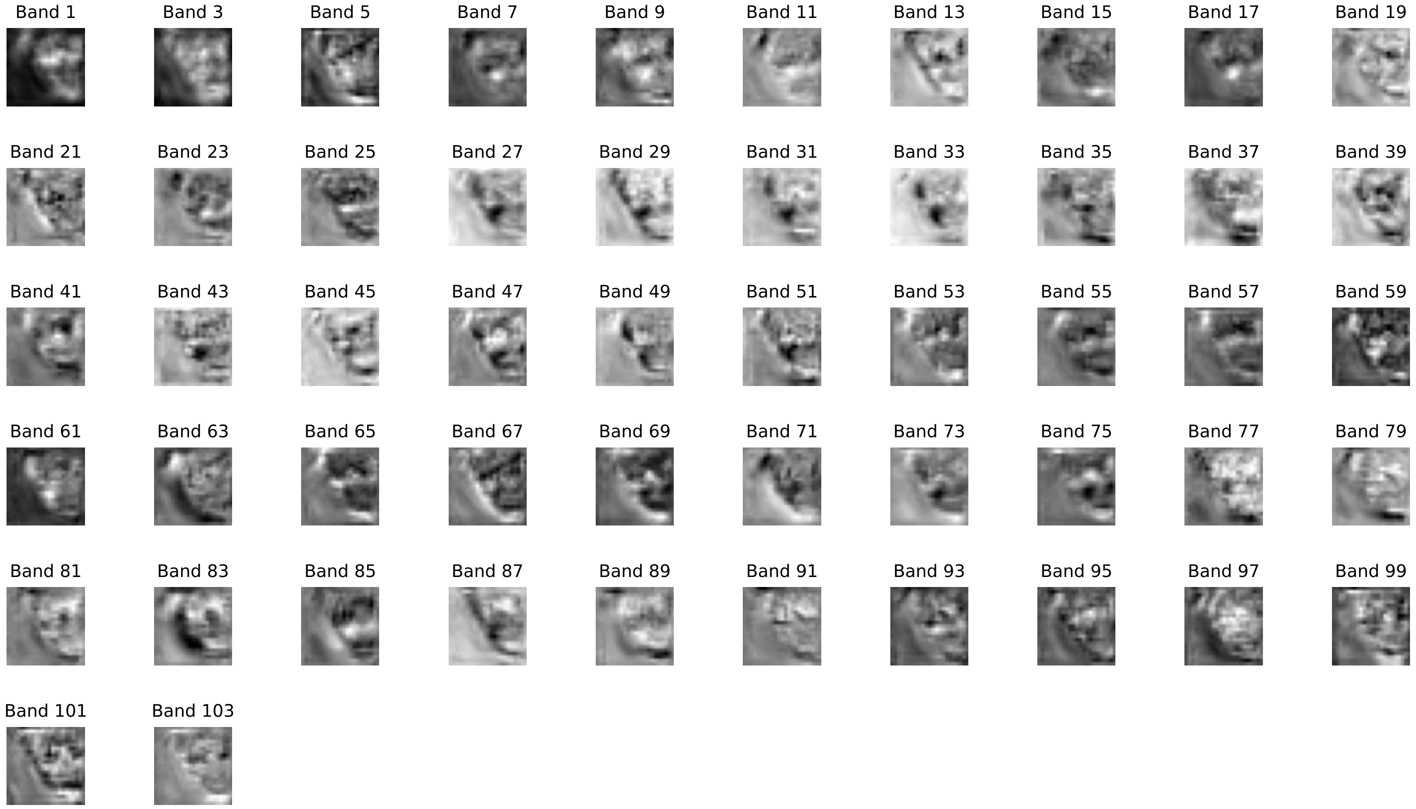


Figure B.3 Guided Backpropagation results for “Trees” class across odd-numbered bands.

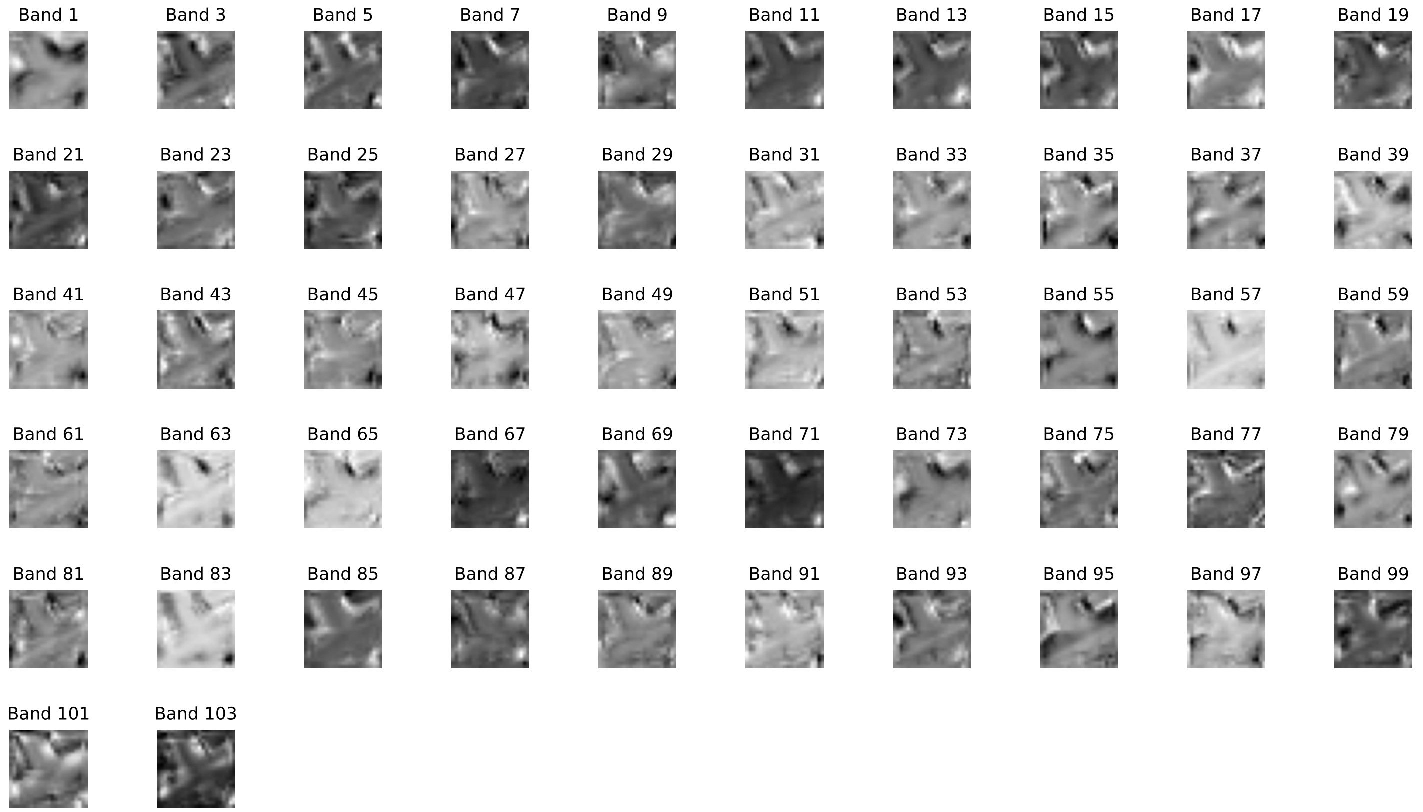


Figure B.4 Guided Backpropagation results for “Painted metal sheets” class across odd-numbered bands.

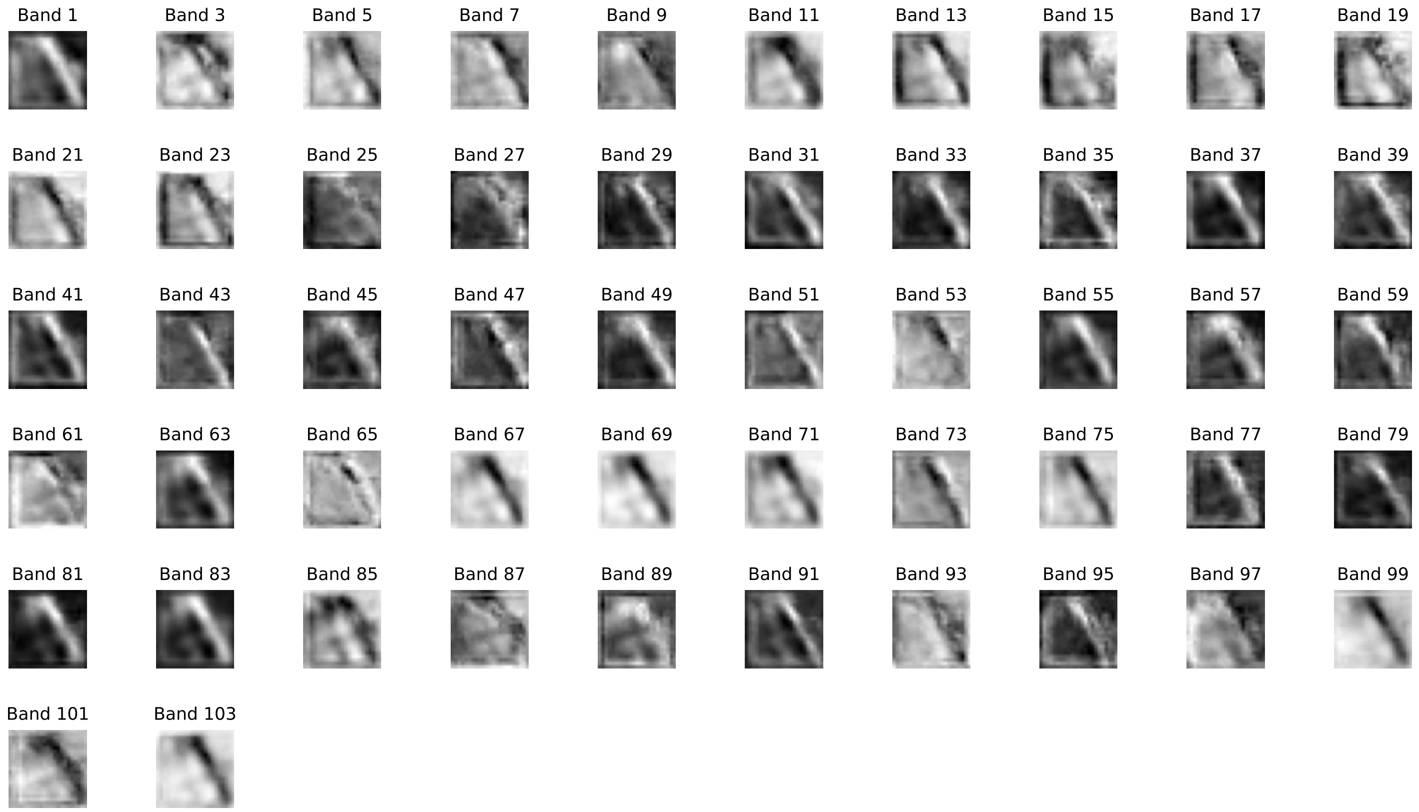


Figure B.5 Guided Backpropagation results for “Bare Soil” class across odd-numbered bands.

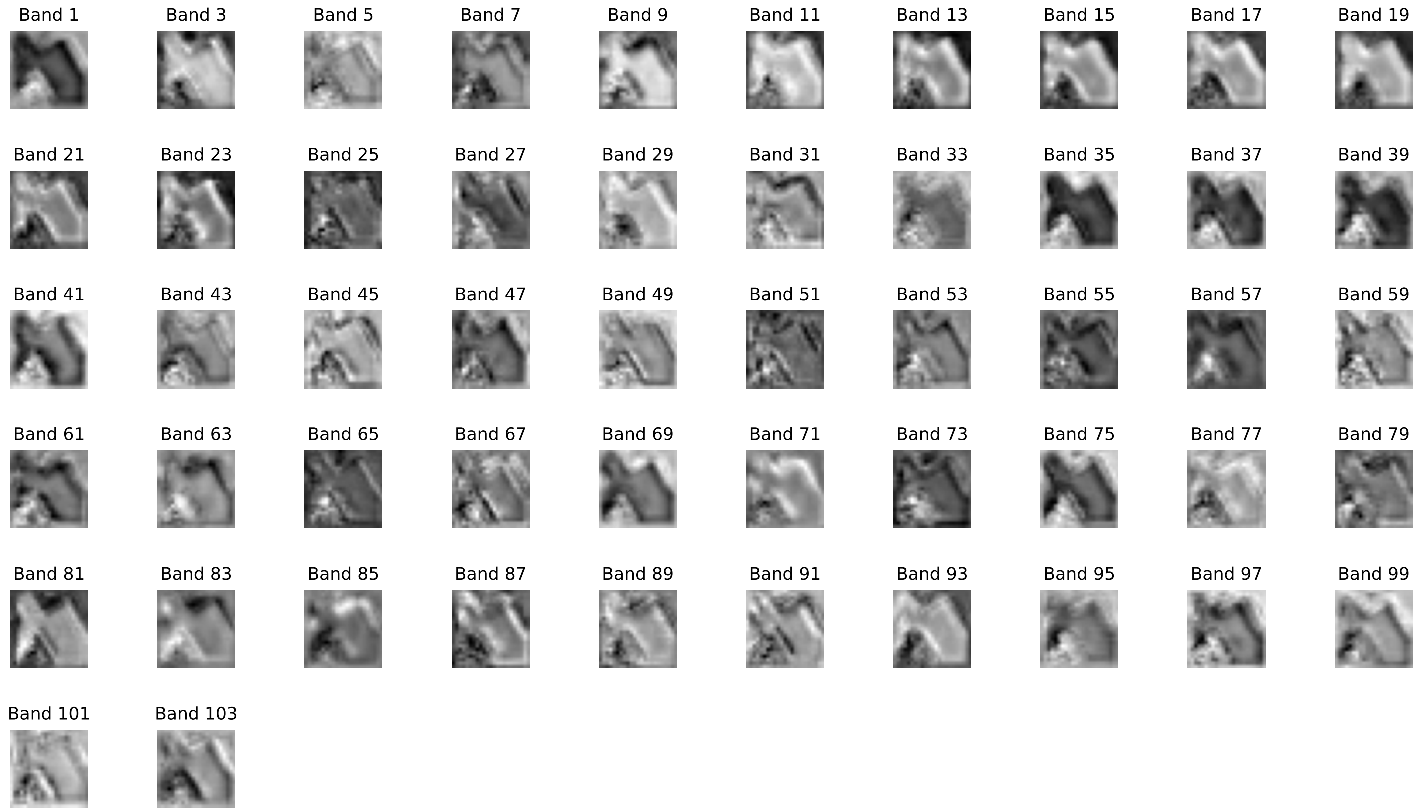


Figure B.6 Guided Backpropagation results for “Bitumen” class across odd-numbered bands.

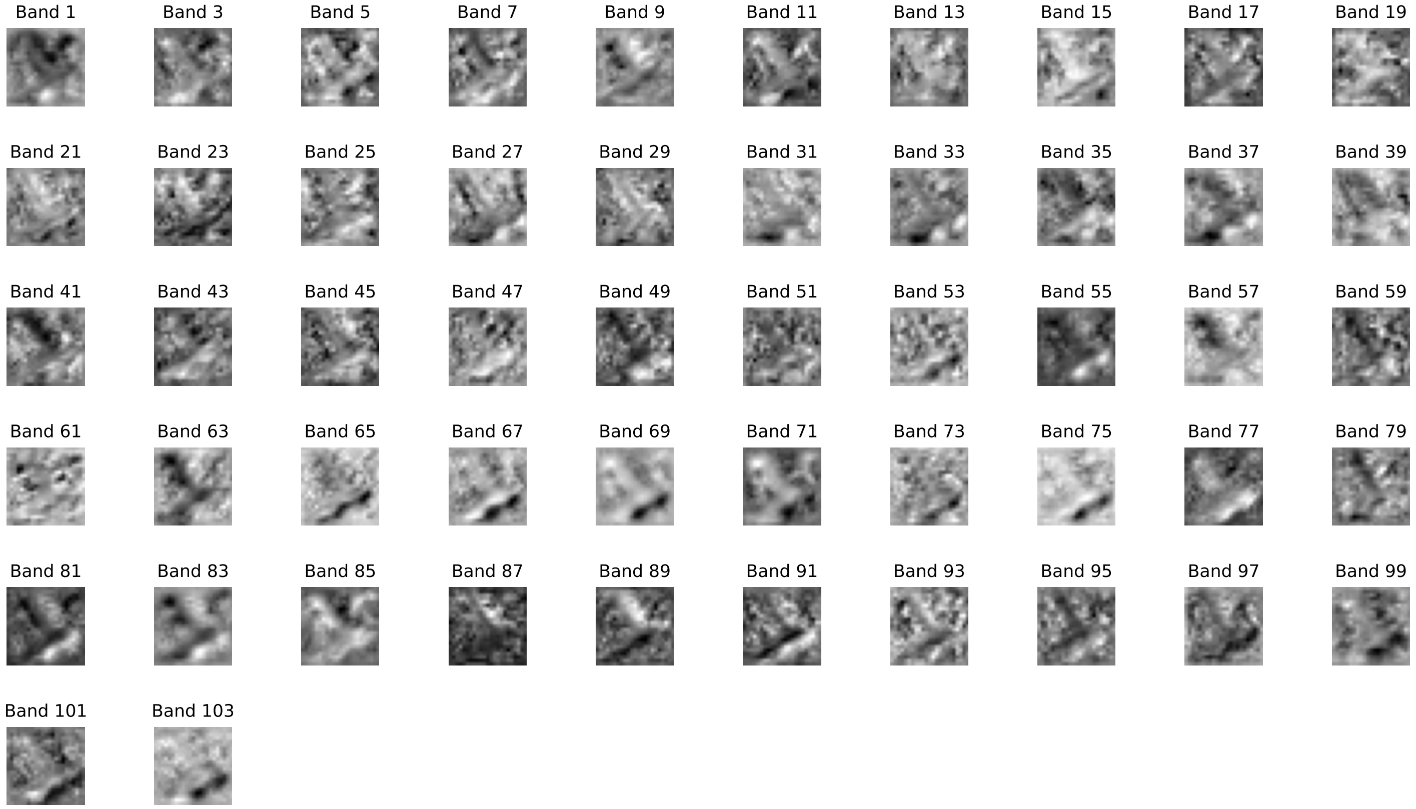


Figure B.7 Guided Backpropagation results for “Self-Blocking Bricks” class across odd-numbered bands.

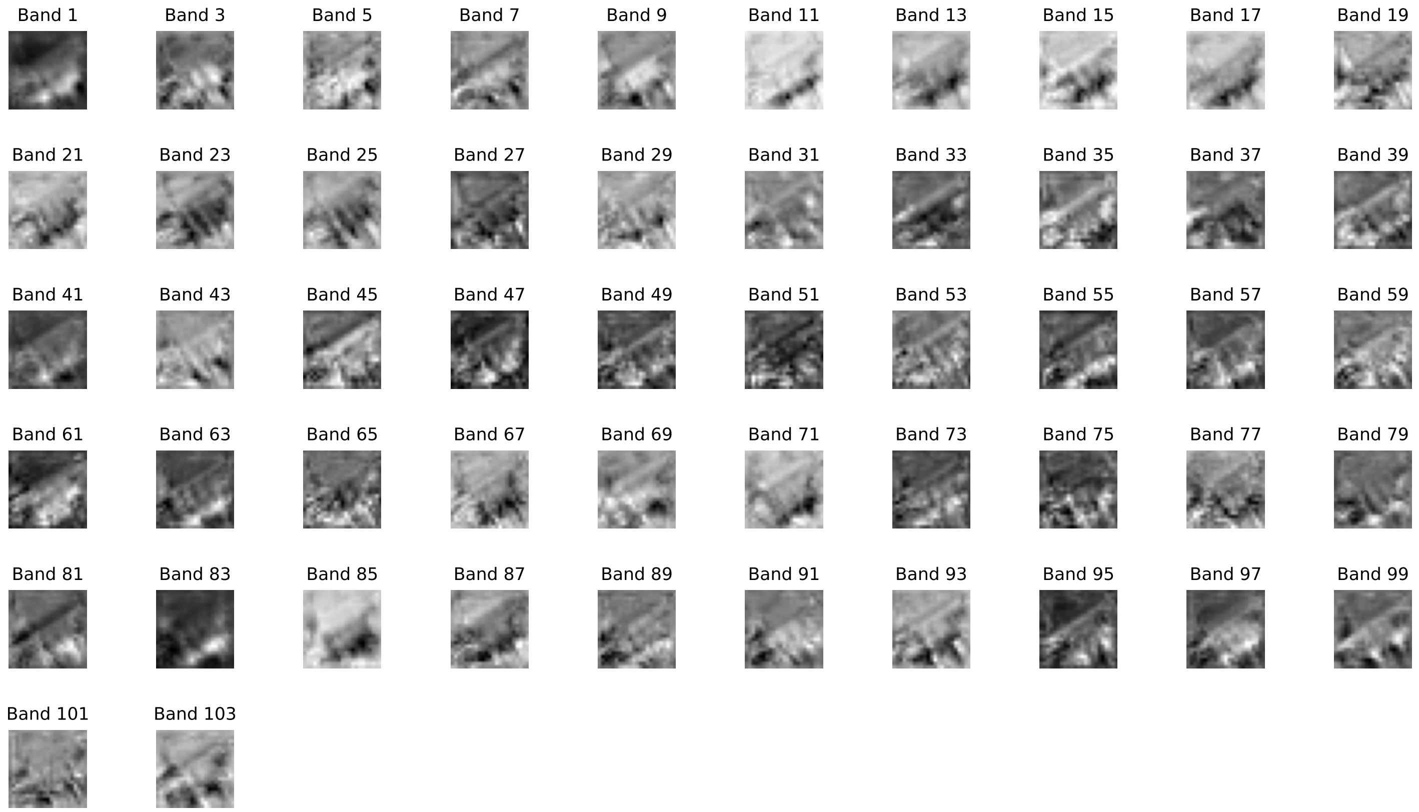


Figure B.8 Guided Backpropagation results for “Shadows” class across odd-numbered bands.

APPENDIX C

List of Papers Published Based on This Thesis

Turan, D. E., Aptoula, E., Ertürk, A., & Taskin, G. (2023). Interpreting hyperspectral remote sensing image classification methods via explainable artificial intelligence. In IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium, (pp. 5950–5953). (Turan, Aptoula, Ertürk & Taskin, 2023)