Computational Science and Engineering
(International Master's Program)

Technische Universität München

Master's Thesis

Detection of Fake Multispectral Remote Sensing Images Using Spectrum-Based Deep Learning and Subspace Learning Techniques

Hilal YILDIZ

Computational Science and Engineering
(International Master's Program)

Technische Universität München

Master's Thesis

Detection of Fake Multispectral Remote Sensing Images Using Spectrum-Based Deep Learning and Subspace Learning Techniques

Author:	Hilal YILDIZ
1st examiner:	Prof. Dr. -Ing. habil. Xiaoxiang Zhu
2nd examiner:	Dr. Daniela Espinoza Molina
3rd examiner:	Dr. Luca Chiarabini
Submission Date:	April 14th, 2026

I hereby declare that this thesis is entirely the result of my own work except where otherwise indicated. I have only used the resources given in the list of references.

April 14th, 2026 Hilal YILDIZ

Acknowledgments

I would like to express my deepest gratitude to my mother, Melek Yıldız, who has been the strongest pillar in my life. Her unwavering belief in me, both emotionally and materially, has carried me through every challenge and made this journey possible. Her endless patience, unconditional love, and countless sacrifices have shaped who I am today. I am forever grateful for her constant encouragement and strength. I would also like to sincerely thank my brother, Mustafa Yıldız, for always being by my side. His support, presence, and quiet strength have meant more to me than words can express. I am truly thankful to my supervisor, Daniela Espinoza Molina, for her kindness, positivity, and continuous support throughout this journey. Her warm and encouraging attitude created a motivating and inspiring environment, and her guidance has been invaluable to the completion of this work. I would also like to sincerely thank my supervisor, Luca Chiarabini, for his insightful feedback, support, and guidance during my time at DLR. His encouragement and willingness to help have played an important role in shaping this thesis. They have been far more than supervisors, offering not only academic guidance but also genuine encouragement and understanding. Finally, I would like to honour the memory of my late father, Kadir Yıldız. Although he is no longer with me, his strength, values, and belief in the power of education continue to guide and inspire me every day. His presence lives on in everything I achieve.
Although this may appear to be a small step, it holds deep personal meaning and marks a significant milestone in an ongoing journey of becoming -shaped by both challenges and hope- with the aim of contributing to the greater good of society and humanity.

Abstract

The rapid advancement of image generation and manipulation techniques has significantly increased the prevalence of highly realistic fake imagery, posing critical challenges for applications that rely on the integrity of remote sensing data. Detecting such manipulations is particularly difficult due to the heterogeneous nature of forgery mechanisms, which introduce fundamentally different and often non-overlapping artefacts.

This thesis investigates the detection of fake remote sensing images from a representation-driven perspective. Challenging the conventional assumption that all manipulated images share common detectable patterns, this work demonstrates that different manipulation types—specifically GAN-generated images and copy-move forgery (CMF)—exhibit distinct characteristics that require fundamentally different feature representations. Through systematic analysis using frequency-domain (FFT), wavelet-based, and data-driven decomposition methods, it is shown that detection performance is strongly dependent on the alignment between feature representation and manipulation characteristics.

To explore this dependency, multiple models are employed, including representation-specific ResNet architectures and the interpretable Geo-DefakeHop framework. Experimental results reveal that detection models do not fail randomly; instead, their failures are structured and directly linked to the type of representation they employ. Models trained on specific manipulation types exhibit limited generalisation due to a fundamental mismatch between the learned features and the underlying artefacts.

To address this limitation, a representation-aware hierarchical ensemble framework is proposed. The framework integrates multiple specialised models through a conditional decision mechanism that adaptively selects and combines representations based on input characteristics. Unlike conventional ensemble methods, this approach explicitly accounts for representation compatibility, enabling the system to leverage complementary strengths while reducing conflicting predictions. As a result, the proposed framework achieves more robust and balanced performance across diverse manipulation types without requiring prior knowledge of the forgery type.

Overall, this work establishes that fake image detection is inherently a multi-characteristic and representation-dependent problem. It provides both empirical evidence and a principled framework for transitioning from single-model detection to adaptive, representation-aware systems, offering improved robustness, interpretability, and generalisation in remote sensing image forensics.

Contents

Part I.

Introduction

1. Introduction

1.1. Introduction

The rapid advancement of image generation and manipulation techniques has led to a significant increase in the availability of highly realistic fake imagery. This development poses serious risks in domains where visual data integrity is critical, such as remote sensing, security, and digital forensics. In these applications, incorrect interpretation of manipulated imagery may lead to faulty decision-making with potentially severe consequences. In remote sensing, where imagery serves as a primary source for environmental monitoring, border security, and strategic intelligence, the emergence of highly realistic, manipulated geospatial imagery could pose significant geopolitical risks or compromise disaster response efforts. Unlike natural images, multispectral, high-resolution remote sensing images demand a level of integrity that current single-domain detectors cannot guarantee.

Despite substantial progress in fake image detection, existing approaches largely rely on a single model or a fixed feature representation. These methods implicitly assume that all types of manipulated images share common detectable characteristics. However, this assumption does not hold in practice.

Different manipulation techniques introduce fundamentally different artefacts. For instance, GAN-generated images often exhibit subtle inconsistencies in the frequency domain due to the underlying generation process, whereas copy-move forgery (CMF) produces spatially localised duplications that preserve global image statistics. As a result, detection performance becomes highly dependent on the compatibility between the chosen feature representation and the underlying manipulation characteristics.

This thesis is built on the central hypothesis that *fake image detection is strongly dependent on the choice of feature representation*. In other words, detection models do not fail randomly; instead, their failures are systematic and directly linked to the type of features they are designed to capture. Models that rely on frequency-domain representations perform well on GAN-generated images but fail to detect spatial manipulations such as CMF. Conversely, spatially focused models succeed in detecting CMF but struggle with GAN-based artefacts.

To investigate this hypothesis, this work conducts a systematic analysis of multiple representation strategies, including frequency-domain, wavelet-based, and data-driven feature decomposition methods. The analysis reveals that each model captures only a subset of manipulation characteristics, leading to complementary strengths and weaknesses. This observation highlights a fundamental limitation of current approaches and motivates the need for adaptive detection strategies. The primary goal of this thesis is to bridge the gap

between specific manipulation types and their optimal feature representations. By systematically deconstructing the failure modes of current models, this work aims to transition from universal yet limited detection to an adaptive, robust hierarchical framework that operates without prior knowledge of the manipulation type.

Motivated by these findings, this thesis proposes a representation-aware hierarchical ensemble framework for fake image detection. Instead of relying on a single model, the proposed approach introduces a conditional decision mechanism that adaptively selects and combines models based on the characteristics of the input image. This enables the system to leverage complementary representations and improve robustness across diverse manipulation types without requiring prior knowledge of the forgery type.

1.2. Thesis Contribution

The main contributions of this thesis are summarised as follows:

- **Development and Analysis of Two Detection Approaches:** Two complementary methods for fake remote sensing image detection are implemented and analysed, namely (i) ResNet-based models operating on frequency representations, and (ii) the Geo-DefakeHop framework for interpretable, channel-wise analysis.
- **Proposal of a Representation-Aware Hierarchical Ensemble:** An ensemble framework is proposed to combine complementary models and improve robustness across different manipulation types.
- **Framework Implementation and Reproducibility:** All proposed methods and experimental pipelines are implemented and integrated into a unified framework, and contributed to a GitHub repository.
- **Systematic Documentation and Analysis of Experiments:** All conducted experiments are thoroughly documented and analysed across multiple datasets, supported by interpretability techniques such as Grad-CAM and patch-level visualisations to explain model behaviour.

Importantly, this work not only improves detection performance but also provides insight into why models succeed or fail under different manipulation scenarios.

Overall, this work challenges the conventional assumption that a single model can robustly detect all types of fake images. Instead, it shows that fake image detection should be treated as a multi-characteristic problem, requiring adaptive, representation-aware solutions.

The remainder of this thesis is structured as follows: **Chapter 2** reviews the literature on remote sensing analysis and forgery detection. **Chapter 3** describes the datasets and their spectral characteristics. The proposed methodology, including ResNet-based and Geo-DefakeHop frameworks, is detailed in **Chapter 4**. **Chapter 5** presents the experimental

results, interpretability analyses, and cross-dataset evaluations. Finally, **Chapter 7** summarises the key findings and outlines future research directions. Together, these chapters provide a comprehensive analysis of representation-dependent behaviour and motivate the proposed ensemble framework.

Part II.

Literature Review

2. Literature Review

2.1. Challenges in Multispectral Remote Sensing Image Forensics

Remote sensing (RS) images play a critical role in applications such as environmental monitoring, urban planning, and security analysis. Due to their high spatial resolution, multiband structure, and large coverage areas, RS images contain complex patterns and rich semantic information. However, these same properties also increase their vulnerability to manipulation and make reliable analysis significantly more challenging [14].

Unlike natural images, remote sensing imagery introduces additional complexities for forgery detection. These include significant variations in scale and viewpoint, heterogeneous spectral characteristics, and numerous small objects distributed over wide spatial regions[38]. Such properties fundamentally differ from those of conventional natural image datasets, leading to a mismatch between the assumptions underlying traditional image forensics methods and the characteristics of RS data. As a result, many existing approaches struggle to generalise effectively when applied to remote sensing imagery[4].

Furthermore, the limited availability of large-scale RS forgery datasets restricts the effectiveness of data-driven methods and further exacerbates generalisation issues [4]. This scarcity not only affects model performance but also limits the ability to learn robust and transferable manipulation features.

Overall, the challenges in remote sensing image analysis primarily stem from the complexity and heterogeneity of the data itself. This data-centric difficulty is further amplified by the diversity of manipulation mechanisms, which introduce fundamentally different types of artefacts[6].

2.2. Fake Image Generation Methods

The diversity of manipulation mechanisms plays a central role in the difficulty of detecting fake images.

Fake image generation methods can be broadly categorised into *learning-based manipulation techniques* and *region-based manipulation techniques*, which differ not only in how fake content is produced but also in the nature of the artefacts they introduce[35]. Among these categories, GAN-based methods represent the most prominent class of learning-based manipulations, while copy-move forgery (CMF) is a representative example of region-based manipulation. These two paradigms are particularly important as they introduce funda-

mentally different artefact characteristics, making them suitable for analysing the challenges of fake image detection.

Generative Adversarial Networks (GANs) are among the most influential frameworks for synthetic image generation. Models such as CycleGAN enable unpaired image-to-image translation through cycle-consistency constraints, allowing realistic transformations between domains[4]. Despite their high visual quality, GAN-generated images often exhibit subtle inconsistencies in texture and high-frequency details. Even advanced models such as StyleGAN2, which reduce visible artefacts and improve visual realism, do not fully capture natural frequency distributions, particularly in high-frequency components, resulting in detectable spectral discrepancies [22].

These inconsistencies are largely attributed to up-sampling operations in generative networks, which introduce systematic distortions in the frequency domain [36]. Additionally, GANs exhibit a frequency bias, prioritising low-frequency content while underrepresenting fine-grained details. As a result, GAN-generated images are primarily characterised by *global frequency-domain artefacts*.

In contrast, **copy-move forgery (CMF)** is a classical manipulation technique in which a region of an image is duplicated within the same image. Since the copied region shares identical colour, texture, and noise properties with the original content, CMF does not introduce new synthetic information. Instead, it alters the image's spatial structure, making detection inherently challenging [2].

Unlike GAN-based methods, CMF preserves global image statistics while introducing *localised structural inconsistencies* in the form of duplicated or transformed regions. As a result, CMF detection relies on identifying intra-image similarity rather than global statistical deviations[8]. Overall, GAN-based generation and CMF represent fundamentally different manipulation paradigms, introducing largely distinct artefact characteristics.

2.3. Fake Image Detection Methods

2.3.1. GAN-generated Image Detection

The increasing realism of GAN-generated images has motivated extensive research on detection methods. Early approaches primarily focus on spatial artefacts, such as colour inconsistencies, checkerboard patterns, and structural irregularities introduced during image generation. While effective for earlier GAN models, these artefacts have become increasingly less reliable as modern architectures produce more visually realistic outputs[30].

To address this limitation, more recent approaches shift towards frequency-domain analysis. It has been shown that GAN-generated images exhibit systematic discrepancies in high-frequency components, which can serve as reliable forensic cues. These spectral artefacts arise from limitations of generative models, particularly their inability to accurately reproduce natural image statistics. As a result, frequency-based representations have emerged as a powerful tool for detecting GAN-generated content[12].

Recent studies have specifically tailored these frequency-domain insights to the unique challenges of remote sensing data. For instance, a neural network-based spectral detector was investigated, designed to identify artefacts produced by the upsampling blocks inherent in most generative adversarial networks (GANs). Their work emphasises that although many detectors are tested on fully forged images, detecting partial inpainting in RS imagery—where only specific regions are counterfeited—remains a more complex and practically relevant challenge. By utilising a ResNet-34 architecture trained on amplitude-spectrum-derived spectral features, their approach demonstrated strong performance in detecting localised forgeries produced by advanced generators, such as those employing dedicated edge-learning subnets. This further underscores the potency of spectral analysis in bridging the gap between general image forensics and the high-resolution requirements of remote sensing analysis[13].

Another important observation is the *spectral bias* of neural networks, where models tend to prioritise learning low-frequency components over high-frequency details [29]. This bias further amplifies the distinction between real and generated images, as GAN-generated images often fail to capture fine-grained high-frequency structures. Consequently, detection methods that explicitly leverage frequency-domain information can better exploit these inconsistencies.

In addition to frequency-based approaches, several methods focus on identifying model-specific fingerprints embedded in generated images. These fingerprints arise from architectural design choices and training dynamics, enabling not only detection but also attribution to specific generative models [25].

Although certain artefact patterns may be shared across some GAN models, their characteristics often vary depending on the underlying architecture and generation process. As a result, their performance may degrade when applied to images generated by unseen or more advanced models. This lack of generalisation suggests that many approaches implicitly rely on specific artefact patterns rather than learning robust, manipulation-invariant representations[13].

Furthermore, most detection methods are designed under the assumption that GAN-generated images share common and, detectable characteristics[41]. However, as discussed in the previous section, manipulation artefacts vary significantly across different generation mechanisms and model architectures. This mismatch between assumed and actual artefact distributions limits the effectiveness of single-representation approaches.

Overall, these findings indicate that while frequency-based and fingerprint-based methods are effective within specific settings, they have limited ability to generalise across diverse generative models. This limitation suggests that GAN detection appears to be representation-dependent, motivating the exploration of more adaptive and complementary detection strategies.

2.3.2. Copy-Move Forgery Detection Methods

Copy-move forgery detection (CMFD) is a fundamental problem in digital image forensics, particularly challenging due to the high similarity between the source and duplicated regions[21]. Unlike generative manipulations, CMF does not introduce new content but instead reuses existing image regions, preserving global image statistics while altering spatial structure[28]. This makes detection inherently difficult, as the manipulation is often visually indistinguishable from authentic content.

Existing CMFD methods can be broadly categorised into three groups: keypoint-based, block-based, and deep learning-based approaches, each targeting different aspects of similarity detection[34].

Keypoint-based Methods

Keypoint-based methods extract distinctive local features, such as SIFT or SURF, and perform matching across the image to identify duplicated regions[15]. These approaches are computationally efficient and robust to geometric transformations such as rotation and scaling. However, their effectiveness is limited in smooth or low-texture regions where distinctive keypoints are scarce[16]. As a result, they tend to struggle to detect forgeries in homogeneous areas, which is common in remote sensing imagery[40].

Block-based Methods

Block-based methods divide the image into overlapping or non-overlapping regions and extract features such as DCT coefficients, PCA components, or wavelet representations[21]. These methods are better suited for detecting duplicated regions in low-texture areas and can handle certain post-processing operations, such as compression and noise reduction. However, block-based approaches suffer from high computational complexity due to exhaustive matching and are sensitive to geometric transformations such as scaling and rotation[18]. Furthermore, their reliance on local similarity makes them vulnerable to false positives when naturally repetitive patterns are present.

Deep Learning-based Methods

Deep learning approaches have significantly improved CMFD performance by enabling automatic feature extraction and end-to-end learning. Convolutional Neural Networks (CNNs), segmentation-based architectures, and multi-scale models are widely used for detecting and localising forged regions[1].

Recent methods formulate CMFD as a dense prediction or segmentation problem, allowing pixel-level localisation of tampered regions[37][33]. Advanced architectures incorporate multi-branch designs, attention mechanisms, and feature fusion strategies to enhance robustness and detection accuracy [28].

Despite these advances, deep learning-based CMFD methods exhibit several limitations. They require large-scale annotated datasets, which are often scarce in remote sensing scenarios, and their performance is highly dependent on the characteristics of the training data[42]. Moreover, these models tend to learn similarity patterns specific to particular datasets, leading to limited generalisation across different manipulation settings[44].

More fundamentally, many CMFD approaches rely on detecting intra-image similarity as the primary cue for forgery. While effective for identifying duplicated regions, this strategy introduces ambiguity, as natural images often contain repetitive structures that resemble manipulated patterns[44]. As a result, models may produce false positives in real images or struggle to detect subtle manipulations when similarity cues are weak.

Overall, these limitations indicate that CMF detection is primarily driven by spatial similarity rather than global statistical inconsistencies. This distinguishes it from GAN-generated image detection and suggests that detection methods benefit from relying on representation strategies that emphasise local structural relationships[14][35]. Consequently, approaches designed for frequency-based artefact detection may have limited transferability to CMF, further highlighting the need for representation-aware and complementary detection frameworks[11].

2.4. Limitations and Challenges

Despite significant progress, fake image detection remains a challenging problem due to several fundamental limitations that are closely tied to the nature of manipulation artefacts and feature representations.

- **Limited Generalisation:** Detection models are typically trained on specific manipulation types and tend to learn artefact patterns that are characteristic of the training data[13]. As a result, their performance may degrade when applied to unseen manipulation types or images generated by different models. This suggests that many approaches rely on manipulation-specific cues rather than learning generalisable representations.
- **Data Dependency:** Deep learning-based methods require large-scale annotated datasets to achieve strong performance[42]. However, in specialised domains such as remote sensing, such datasets are limited in size and diversity. This restricts models' ability to learn robust, transferable features, further exacerbating generalisation issues.
- **Frequency Bias:** Neural networks exhibit a well-known bias towards low-frequency components, often neglecting high-frequency details [41]. While high-frequency artefacts are critical for detecting GAN-generated images, this bias can limit the effectiveness of standard deep learning approaches unless explicitly addressed.
- **Heterogeneity of Forgeries:** Different manipulation techniques introduce fundamentally different types of artefacts. GAN-generated images are characterised by

global frequency inconsistencies[41], whereas copy-move forgery relies on localised spatial duplication[44]. These differences suggest that a single feature representation may not be sufficient to effectively capture all types of manipulation.

More importantly, these challenges are not independent but are intrinsically connected through the concept of representation. The limited ability of models to generalise, their sensitivity to data distribution, and their dependence on specific artefact types all suggest a mismatch between the chosen feature representation and the underlying manipulation characteristics. This highlights an important limitation of many existing approaches, which often assume that a single model or representation can effectively capture different types of manipulation artefacts. However, as discussed in the previous sections, this assumption may not always hold in practice [13].

Therefore, effective fake image detection calls for adaptive strategies that explicitly account for the diversity of manipulation mechanisms and the representation-dependent nature of detection[14]. This motivates approaches that combine complementary feature representations and leverage their strengths in a structured manner[35].

2.5. Motivation for Multi-Model Approaches

GAN-generated images can be more effectively detected using frequency-domain representations [41], whereas copy-move forgery (CMF) relies on spatial-similarity analysis [44]. This observation highlights a key limitation of existing approaches: many detection methods rely on a single model and a fixed feature representation, implicitly assuming that all manipulation types share similar detectable patterns. However, as discussed in the previous sections, this assumption may not always hold in practice and can lead to performance degradation when models are applied to heterogeneous manipulation types [25, 13].

To address this limitation, multi-model or ensemble-based frameworks have been proposed to combine complementary detection strategies[28]. By integrating deep learning, frequency-domain analysis, and statistical methods, these systems aim to leverage diverse feature representations and improve robustness across different manipulation scenarios.

However, existing ensemble approaches often combine models in a uniform or parallel manner, without explicitly considering the relationship between feature representations and manipulation characteristics[28]. As a result, they may fail to fully exploit the complementary nature of different models, leading to inconsistent or conflicting predictions.

These limitations suggest the need for more structured and adaptive ensemble strategies[35]. In particular, hierarchical and conditional frameworks enable dynamic model selection based on input characteristics, allowing the system to prioritise the most relevant representation for a given manipulation type.

Therefore, effective fake image detection benefits from not only combining multiple models but also explicitly aligning model selection with manipulation-specific characteristics. This perspective motivates the development of representation-aware multi-model

frameworks, where complementary models are integrated in a structured and adaptive manner with the goal of improving robustness and generalisation.

Part III.

Methodology

3. Datasets

The datasets used in this study are selected for their relevance to state-of-the-art (SOTA) research on fake remote sensing image detection. In particular, the selected datasets are widely used in the literature and serve as benchmarks in recent studies.

This choice ensures that the experimental evaluation aligns with current research practices and enables meaningful comparison with existing approaches. Additionally, the datasets represent different manipulation paradigms, including both generative model-based synthesis and spatial manipulation-based forgery, reflecting the diversity of scenarios addressed in prior work.

Due to the limited availability of publicly accessible remote sensing forgery datasets, a set of representative and widely used datasets is selected to balance diversity and experimental feasibility.

3.1. Selected Datasets

The datasets used in this study include the FSI dataset, the DM-AER dataset, and a copy-move forgery (CMF) dataset derived from the RSCMQA framework. These datasets are selected to reflect different manipulation mechanisms and to enable a comparative analysis of their characteristic artefacts.

3.1.1. The Fake Satellite Imagery Dataset

The FSI dataset[43] consists of 2,032 real and 2,032 fake images generated using CycleGAN. The generated images introduce subtle inconsistencies, particularly in frequency distributions.

As shown in Figure 3.1, the FSI dataset contains both real and CycleGAN-generated satellite images. Although the fake images appear visually realistic, they may contain subtle inconsistencies that are not easily observable in the spatial domain.

3.1.2. The DeepMedia Aerial Deepfake Dataset

Similar to the FSI dataset, the DM-AER dataset[9] contains aerial imagery generated using StyleGAN2 under diverse environmental conditions[9]. To maintain computational feasibility and class balance, a subset of 4,000 images (2,000 real and 2,000 fake) is used.

Figure 3.1.: Example images from the FSI (Fake Satellite Imagery) dataset. (a) Original map tile, (b) corresponding real satellite image, and (c)-(d) CycleGAN-generated fake images incorporating visual patterns from different geographic regions. Adapted from [43].

As shown in Figure 3.2, the DM-AER dataset includes both real and StyleGAN2-generated aerial images. The generated samples closely resemble real images in terms of visual appearance, reflecting the high quality of modern generative models.

3.1.3. Copy-Move Forgery Dataset

In contrast to GAN-based datasets, the CMF dataset is constructed from real remote sensing imagery, where manipulation is introduced by duplicating regions within the same image.

CMF datasets involve duplicating image regions, resulting in manipulated areas that share visual characteristics with the original content. This makes them suitable for evaluating spatially localised representations.

As shown in Figure 3.3, CMF images are created by duplicating regions within the same image. As a result, the duplicated areas preserve similar colour, texture, and structural properties as the original content.

Although these datasets are introduced individually, they can be grouped based on their underlying manipulation mechanisms. Specifically, FSI and DM-AER represent generative model-based manipulations, whereas the CMF dataset represents spatial manipulation-based forgery.

(a) Real (b) Fake (StyleGAN2)

(c) Real (d) Fake (StyleGAN2)

Figure 3.2.: Example images from the DM-AER dataset. Each row presents a real aerial image (left) and a corresponding StyleGAN2-generated fake image (right).

The selection of these datasets is directly motivated by the central hypothesis of this thesis. Each dataset represents a distinct type of manipulation with fundamentally different characteristic properties.

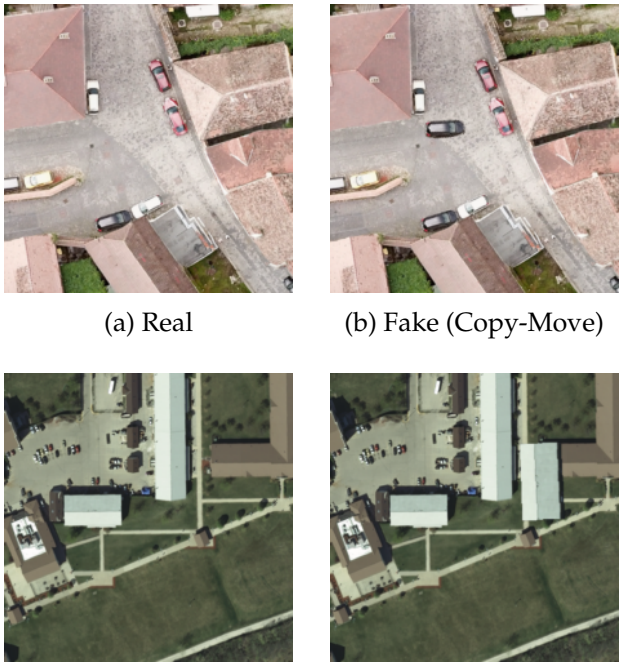
GAN-generated datasets introduce globally distributed artefacts, particularly in the frequency domain, as a result of the generative process. In contrast, CMF datasets preserve global image statistics while introducing localised structural inconsistencies through region duplication.

Although both FSI and DM-AER datasets are GAN-generated, they are produced using different generative models, namely CycleGAN and StyleGAN2, which rely on distinct generation mechanisms. As a result, they exhibit different artefact characteristics despite belonging to the same manipulation category.

Together, these differences enable a systematic investigation of how different feature representations respond to different manipulation characteristics. This design is essential for analysing whether detection models learn generalisable features or remain dependent on specific manipulation types.

Table 3.1 summarises the datasets used in this study.

To enable a fair and controlled comparison, all datasets are balanced with equal numbers of real and fake samples. This reduces potential bias toward a particular class and helps



(a) Real (b) Fake (Copy-Move)

(c) Real (d) Fake (Copy-Move)

Figure 3.3.: Example images from the CMF dataset. Each row presents a real image (left) and a corresponding copy-move manipulated image (right).

Table 3.1.: Summary of datasets used in this study

Dataset	Manipulation Type	Real	Fake	Total
FSI	CycleGAN	2,032	2,032	4,064
DM-AER (subset)	StyleGAN2	2,000	2,000	4,000
CMF (RSCMQA-based)	Copy-Move	2,000	2,000	4,000

ensure that performance differences are primarily attributable to representation compatibility rather than class imbalance.

3.2. Dataset Characteristics

Understanding the differences among manipulation types is important for analysing the role of feature representation in fake image detection. To examine these differences, frequency-domain representations are analysed using the Discrete Cosine Transform (DCT).

Previous studies suggest that GAN-generated images may exhibit irregularities in mid- and high-frequency regions due to the generative process, resulting in non-uniform spectral responses [36]. In contrast, real images typically exhibit smoother, more natural frequency distributions.

As illustrated in Figure 3.4, GAN-generated images tend to exhibit stronger and more irregular energy distributions in high-frequency regions, which may reflect artefacts introduced during image synthesis. In contrast, real images exhibit a more balanced and smoothly decaying frequency distribution.

For CMF images, no significant global deviation is typically observed in the frequency domain. Since manipulated regions originate from the same image, the overall frequency distribution remains largely unchanged, making global frequency-based cues less informative.

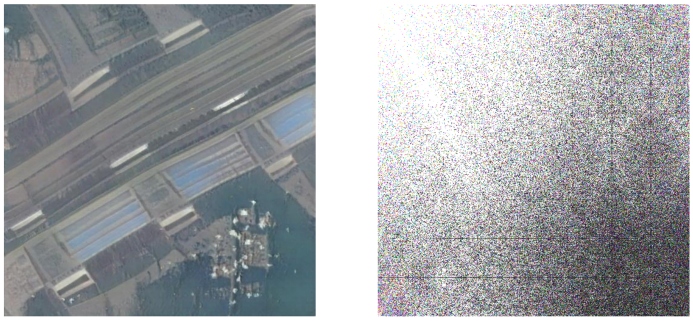
To better capture CMF characteristics, wavelet-based representations are considered. Unlike FFT and DCT, which provide global frequency analysis, the Discrete Wavelet Transform (DWT) captures both spatial and frequency localisation [23]. This has the potential to detect localised structural inconsistencies introduced by region duplication.

These observations suggest that different manipulation types may benefit from different feature representations. While GAN-generated images can be analysed using frequency-domain cues [41], CMF manipulations may require representations that capture spatially localised patterns [26].

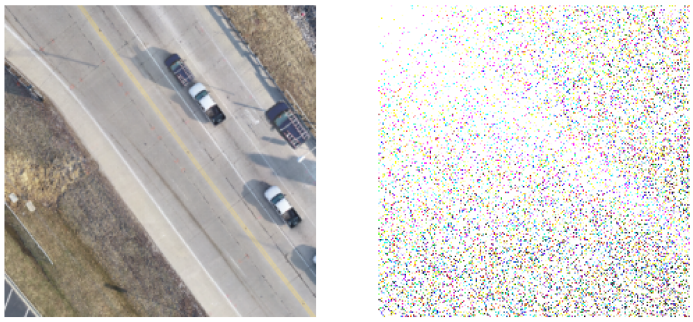
Cross-Dataset Characteristic Differences A comparison between GAN-based and CMF datasets highlights key differences in the nature of manipulation artefacts:

- **Global vs Local Artefacts:** GAN-generated images tend to introduce globally distributed artefacts [41], whereas CMF manipulations result in localised inconsistencies [21].
- **Frequency vs Spatial Patterns:** GAN artefacts are often reflected in the frequency domain[41], while CMF artefacts are characterised by spatial duplication and structural similarity[44].
- **Statistical vs Structural Irregularities:** GAN-generated images may deviate from natural image statistics[36], whereas CMF images tend to preserve global statistics

(a) Real image and its DCT



(b) GAN-generated fake image and its DCT



(c) CMF fake image and its DCT

Figure 3.4.: Comparison of spatial images and their corresponding DCT representations for real, GAN-generated, and CMF images.

but alter structural consistency[21].

These differences suggest that the effectiveness of a detection model may depend on the compatibility between the chosen feature representation and the underlying manipulation characteristics.

As a result, models designed to capture global frequency inconsistencies may be more suitable for GAN-generated images, while being less effective for detecting spatially localised manipulations such as CMF [11]. Conversely, models that focus on spatial similarity may be better suited to CMF detection but may not generalise well to GAN-based artefacts [35].

These observations motivate the use of multiple, representation-specific models, which are further explored in the following chapters.

3.3. Dataset Preparation

Remote sensing (RS) images are inherently multispectral and may contain multiple spectral bands beyond the visible RGB channels. However, in this study, only RGB bands are used. This choice is made to ensure compatibility with standard deep learning architectures and pre-trained models (e.g., ImageNet-based ResNet), which are designed for three-channel inputs. While multispectral information can provide additional discriminative features, the focus of this work is on analysing representation-dependent behaviour under a controlled and widely adopted RGB setting.

All images are resized to a fixed resolution of 224×224 pixels to ensure consistency and compatibility with the ResNet architecture. Multiple input representations are considered to reflect different aspects of the data. FFT is applied to obtain frequency-domain representations, while DWT provides multi-scale spatial-frequency information. In addition, spatial-domain inputs are retained to preserve structural characteristics.

An 80%–20% train-test split is employed, and all datasets are balanced to support a fair evaluation. Despite this design, certain limitations remain. The DM-AER dataset is used as a subset, potentially limiting variability. Additionally, CMF datasets may not fully capture complex real-world manipulations. These limitations are taken into account when interpreting the experimental results and provide directions for future work.

4. Methodology

This chapter addresses the first contribution of this thesis by presenting the development of specialised detection methods and establishing a structured framework for evaluating representation dependency in remote sensing imagery.

This chapter introduces a representation-oriented methodology for detecting manipulated remote sensing images, with an emphasis on examining how different representation strategies relate to observable artefact characteristics.

As discussed in Section 2, manipulation types differ in the way they affect image structure and statistics. These differences motivate the need to consider multiple perspectives when designing detection approaches. Rather than relying on a single representation, this work explores how alternative representations influence detection performance across varying manipulation conditions.

Accordingly, the methodology is structured to enable a comparative analysis of representation strategies. Instead of focusing solely on maximising detection accuracy, the approach aims to provide insight into how models respond to different forms of input transformation and the types of artefacts they can capture.

To support this analysis, a multi-model framework is employed in which each model operates on a specific representation domain. ResNet-based models[41] are used to learn features from frequency and wavelet-transformed inputs, while the Geo-DefakeHop model[4] is incorporated to offer an interpretable perspective on frequency-related characteristics.

Together, these components form a structured experimental setup that facilitates the examination of representation-dependent behaviour. This design allows for a more nuanced understanding of how detection performance varies across manipulation types, without assuming that a single representation is universally sufficient.

4.1. Model Selection and Design Rationale

To investigate how representation strategies influence detection behaviour, two complementary modelling approaches are considered.

ResNet is employed as a deep learning backbone due to its ability to learn hierarchical, distributed feature representations [41]. In this study, it is used as a controlled feature extractor operating on transformed inputs, allowing the analysis of representation-specific learning behaviour. By applying FFT and wavelet transformations, the input is explicitly mapped into feature domains that emphasise different aspects of manipulation artefacts.

In addition to explicit representations, representation strategies are also explored implicitly, using data-driven methods. Geo-DefakeHop[4] is selected as an interpretable statistical framework that decomposes images into frequency-selective channels. Unlike deep learning models, it enables direct examination of how individual components contribute to detection, making it suitable for analysing both representation behaviour and performance.

Together, these models enable a structured comparison between explicit and implicit representation strategies. While ResNet provides high-capacity learning on transformed inputs, Geo-DefakeHop offers an interpretable, data-driven decomposition, allowing complementary perspectives on the detection problem.

4.2. ResNet-Based Detection with Feature-Specific Representations

The proposed framework employs two ResNet-34-based models that operate on different input representations[41]. While the underlying architecture remains unchanged, the input transformation and preprocessing pipeline are adapted to capture distinct types of manipulation artefacts.

For both variants, the standard ResNet architecture is modified for binary classification by replacing the final fully connected layer with a two-class output (real vs fake)[41]. The input layer is adjusted according to the representation: the FFT-based model processes frequency-domain magnitude spectra, while the wavelet-based model operates on multi-channel wavelet sub-band features. This design ensures that observed performance differences are primarily attributable to the choice of representation rather than architectural variation.

The preprocessing pipeline plays an important role in shaping the learning process. By transforming the input into alternative feature domains, the model is guided towards learning representation-driven patterns, such as frequency irregularities or localised structural variations. This enables a controlled comparison of how different input representations influence detection behaviour.

The FFT-based model is initialised with ImageNet pretraining, whereas the wavelet-based model is trained from scratch, as wavelet representations differ from natural image statistics typically encountered during pretraining[39][32].

4.2.1. FFT-Based ResNet for GAN-Generated Images

The input images are first resized to $224 \times 224 \times 3$. The FFT is then applied independently to each RGB channel, producing channel-wise frequency representations. The resulting spectra are converted to magnitude representations and normalised before being used as model input.

A high-pass filtering strategy is applied to suppress low-frequency components and emphasise higher-frequency regions [41]. This transformation encourages the model to focus on patterns that are less prominent in the spatial domain. Such frequency-based representations have been shown to capture artefacts associated with generative processes, particularly those arising from upsampling operations and synthesis inconsistencies [27, 12]. While these artefacts may not be easily observable in the spatial domain, they can become more distinguishable after transformation into the frequency domain.

The resulting representations are used as input to the ResNet model for binary classification (real vs fake). The model is trained and evaluated on GAN-generated datasets, including the FSI dataset (CycleGAN-generated) and the DM-AER dataset (StyleGAN2-generated).

Table 4.1.: Performance metrics of FFT-ResNet

Class	**Precision**	**Recall**	**F1-score**
Fake	0.99	0.97	0.98
Real	0.97	0.99	0.98
Accuracy		**0.98**	

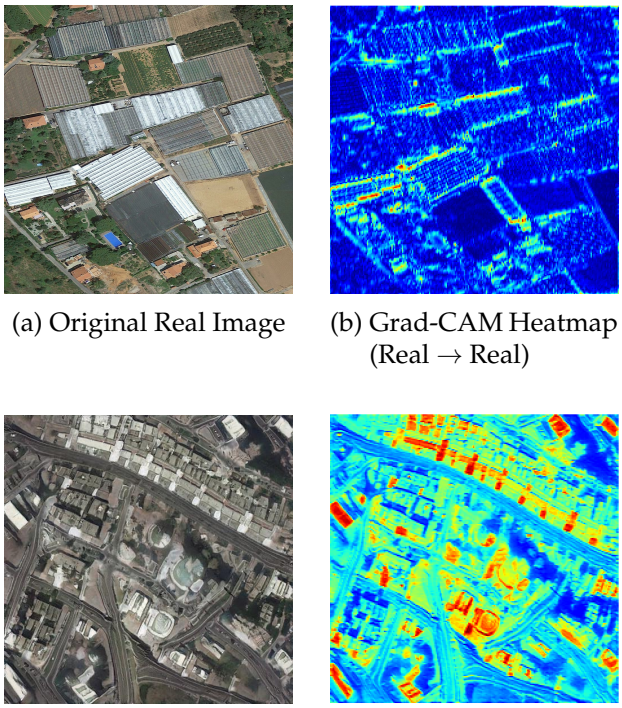
The performance of the proposed model on the FSI and DM-AER datasets is summarised in Table 4.1. The model achieves an overall accuracy of 98.42%, with consistently high precision, recall, and F1 scores for both classes. The recall for fake samples (0.97) suggests that the model is able to detect most manipulated images. The high recall for real samples (0.99) further suggests a low false-positive rate.

These results indicate that frequency-domain representations align with frequency-domain representations that capture artefact-related patterns associated with GAN-generated images, while maintaining strong performance on real samples. This observation is consistent with prior findings that such artefacts may be more distinguishable in the frequency domain[41].

Table 4.2.: Confusion matrix for FFT-ResNet

	Predicted Fake	**Predicted Real**
Actual Fake	185	5
Actual Real	1	189

The confusion matrix, presented in Table 4.2, shows that only 5 fake images are misclassified as real (false negatives) and 1 real image is misclassified as fake (false positive). The low number of false negatives reflects that the model is able to identify most GAN-generated images. Similarly, the low number of false positives indicates that genuine images are rarely misclassified. This balanced error distribution suggests that the model is able to distinguish between real and fake samples under the given experimental conditions.



(a) Original Real Image

(b) Grad-CAM Heatmap (Real → Real)

(c) Original Fake Image

(d) Grad-CAM Heatmap (Fake → Fake)

Figure 4.1.: Grad-CAM visualisations of the FFT-based ResNet model. Top row: real image and corresponding activation map. Bottom row: GAN-generated fake image and activation map. Differences in activation patterns between real and fake samples can be observed, particularly in high-frequency regions.

Grad-CAM visualisations of the model are presented in Figure 4.1. While both real and fake images contain high-frequency components, the activation patterns differ in their spatial organisation.

In real images, high-frequency responses are more irregular and distributed, corresponding to natural edges and textures. In contrast, fake images exhibit more structured, concentrated activation patterns, particularly around repetitive or edge-like regions, consistent with GAN-induced artefacts.

This suggests that the model responds not only to the presence of high-frequency information, but also to differences in its spatial organisation. Such behaviour is consistent with observations that frequency-domain representations can support the distinction between real and GAN-generated images.

4.2.2. Wavelet-Based ResNet for CMF Image Detection

To capture spatially localised patterns associated with copy-move forgery (CMF)[8], a wavelet-based representation is adopted.

In this work, each RGB image is decomposed using the discrete wavelet transform (DWT) into multiple sub-bands (LL, LH, HL, HH), which represent different directional components. These sub-bands are used as multi-channel inputs to the ResNet model, preserving spatial locality while incorporating frequency-related information[24].

This representation allows the model to analyse localised variations within the image and to capture structural patterns associated with region duplication, which are less apparent in global frequency representations[20].

Table 4.3.: Performance metrics of Wavelet-ResNet

Class	Precision	Recall	F1-score
Fake	0.67	0.72	0.70
Real	0.69	0.64	0.66
Accuracy		0.68	

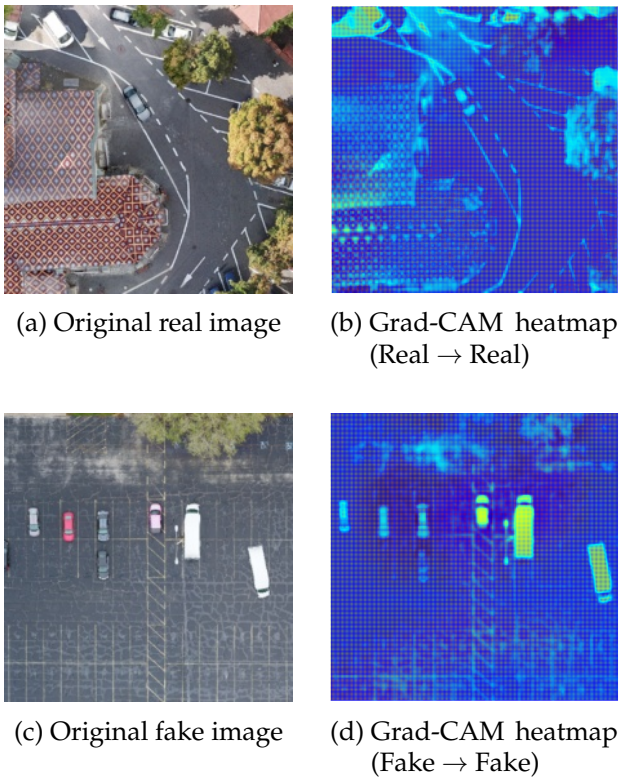
The wavelet-based ResNet model is trained and evaluated on the CMF dataset. The performance is summarised in Table 4.3, where the model achieves an overall accuracy of 68%, with moderate precision and recall for both classes.

Compared to the FFT-based model, the performance is lower under the current experimental setup. This difference may be related to the nature of CMF, where manipulations are characterised by subtle, localised patterns rather than pronounced global artefacts[8, 10]. Such characteristics can make detection more challenging, even when using representations that preserve spatial information.

Table 4.4.: Confusion matrix for Wavelet-ResNet

	Predicted Fake	Predicted Real
Actual Fake	62	24
Actual Real	30	53

The confusion matrix in Table 4.4 shows that 24 fake images are misclassified as real (false negatives) and 30 real images are misclassified as fake (false positives). The relatively high number of false negatives is consistent with the fact that some manipulated regions are not easily distinguishable from natural image structures. Similarly, the presence of false positives indicates that certain patterns in real images may be interpreted as duplicated regions. These observations are consistent with the challenges reported in CMF detection, where distinguishing between natural structural similarity and manipulation-induced duplication can be difficult[8, 10].



(a) Original real image

(b) Grad-CAM heatmap (Real → Real)

(c) Original fake image

(d) Grad-CAM heatmap (Fake → Fake)

Figure 4.2.: Grad-CAM visualisations of the wavelet-based ResNet model. Top row: real image and corresponding activation map. Bottom row: CMF fake image and activation map. Fake images exhibit localised activations over structurally similar regions, while real images exhibit more diffuse responses.

Grad-CAM visualisations for the wavelet-based model are presented in Figure 4.2. The activation maps highlight localised regions in fake images, particularly around small structural elements and repeated patterns.

Compared to the FFT-based model, which emphasises global frequency characteristics, the wavelet-based model produces activations that are more spatially distributed across regions with similar structures. This behaviour suggests that the model is responsive to localised similarities rather than global artefacts.

In contrast, real images exhibit more diffuse, less structured activation patterns, with no consistent regions of emphasis.

These observations are consistent with the idea that the wavelet-based ResNet model is sensitive to structural patterns associated with CMF[8]. However, the comparatively lower performance suggests that detecting such patterns remains challenging under the current experimental conditions. This is consistent with prior work highlighting the difficulty of distinguishing between natural structural similarity and manipulation-induced

duplication[35, 11].

While ResNet-based models achieve strong performance through representation-specific feature learning, they operate as black-box models, providing limited interpretability into which components of the representation contribute to detection.

To further analyse representation behaviour, the next section introduces the Geo-DefakeHop model, which enables an interpretable, channel-wise analysis of frequency components[4].

4.3. Geo-DefakeHop-Based Interpretable Feature Analysis

Geo-DefakeHop is a lightweight, data-driven framework that analyses images using statistical and frequency-aware representations, without relying on end-to-end deep learning. Unlike conventional convolutional models, it explicitly decomposes the input into multiple frequency-selective channels, allowing direct examination of how different components contribute to detection[4].

The method is motivated by prior observations that generative models may exhibit inconsistencies in high-frequency components, even when low-frequency structures are preserved[12]. Geo-DefakeHop leverages this by analysing images across multiple frequency channels.

The model decomposes input images into several subspaces using filter banks[7], where each channel represents a specific spatial-frequency response[4]. Rather than combining features implicitly, Geo-DefakeHop evaluates each channel independently and selects the most discriminative ones based on their classification performance. This enables explicit identification of which frequency components contribute most to detection.

A hierarchical decision strategy is employed. Channel-wise predictions are first obtained at the patch level and then aggregated to form an image-level representation, which is used for final classification[4]. This multi-stage aggregation improves robustness while maintaining interpretability.

By complementing deep learning models with an interpretable statistical framework, Geo-DefakeHop provides additional insight into representation behaviour and enables analysis of how individual frequency components contribute to fake image detection.

4.3.1. Architecture of the Geo-DefakeHop Model

The architecture of the Geo-DefakeHop model is illustrated in Figure 4.3. It consists of a sequence of processing stages, including patch extraction, multi-scale feature decomposition, channel-wise classification, and hierarchical aggregation.

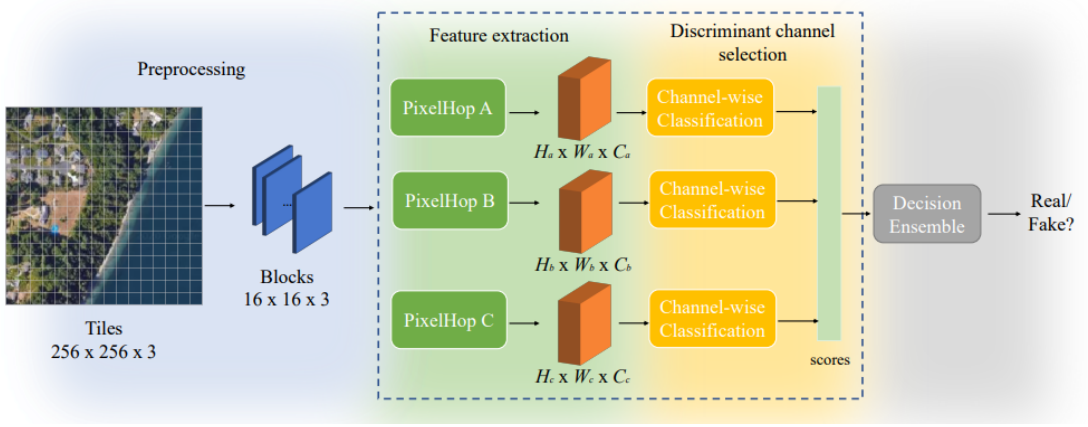


Figure 4.3.: Overview of the Geo-DefakeHop architecture. The input image is divided into patches, followed by multi-scale feature extraction using parallel Saab transforms, channel-wise classification, and hierarchical aggregation for final prediction. Adapted from [4].

Patching and Pre-processing

Input RGB images are resized to 256×256 and partitioned into non-overlapping patches of size $16 \times 16 \times 3$, resulting in 256 patches per image. This patch-based representation facilitates localised analysis of spatial structures[4]. The extracted patches are processed independently during feature extraction and subsequently aggregated to form an image-level representation.

Multi-Scale Feature Extraction via Saab Transform

Feature extraction in Geo-DefakeHop is based on the PixelHop++ framework, which employs a multi-channel Saab transform to analyse spatial-frequency characteristics[7]. This process is implemented in parallel across three branches—Hop A, Hop B, and Hop C—using kernel sizes of (2×2), (3×3), and (4×4), respectively. For each local patch, the Saab transform projects the input into a set of decorrelated basis functions derived from the eigen-decomposition of the local covariance matrix[4].

This transformation decomposes the input into multiple components corresponding to different frequency responses[7]. Unlike fixed convolutional filters, the Saab filters are data-driven and produce decorrelated feature representations aligned with the underlying data distribution.

To maintain a lightweight architecture and retain informative components, a two-stage channel selection mechanism is applied:

1. **Energy-Based Filtering:** The importance of each channel i is evaluated using its nor-

malized eigenvalue:

$$E_i = \frac{\lambda_i}{\sum_j \lambda_j} \tag{4.1}$$

where λ_i denotes the eigenvalue associated with the i^{th} Saab filter. Channels satisfying $E_i > 0.01$ are retained, reducing dimensionality while preserving dominant spectral information[4].

2. **Discriminative Feature Test (DFT):** The remaining channels are evaluated based on their ability to distinguish between real and fake samples[4]. A soft decision score is computed for each channel using class-wise feature distributions, and only channels exhibiting meaningful differences are selected.

The multi-scale design enables analysis across different spatial resolutions[4]. Smaller kernels (Hop A) capture fine-scale variations, while larger kernels (Hop C) capture broader structural patterns[7]. The selected features from all branches are then aggregated for final binary classification.

Channel-Wise Classification and Discriminant Selection

Following feature extraction, each channel is treated as an independent feature component corresponding to a specific frequency response[4]. Geo-DefakeHop evaluates these channels individually rather than combining them into a single representation.

For each channel, patch-level features are used to train a dedicated XGBoost classifier[5], which produces a soft decision score for classification[4].

Channel importance is assessed based on classification performance on a validation set. The channels are ranked accordingly, and only the top-performing ones are retained at each hop[4].

This selection process reduces feature dimensionality while preserving the most discriminative components. The retained channels are then used for subsequent aggregation and final classification.

Patch-Level Feature Aggregation

For each selected channel, the trained XGBoost classifier produces a soft decision score for each patch[4]. These scores are concatenated across channels to form a patch-level representation.

The patch-level representations are then aggregated to obtain an image-level feature vector. The final feature dimension is given by:

$$\text{Feature Dimension} = 256 \times N_{ch}$$

where N_{ch} denotes the number of selected channels[4].

This aggregated representation is used for subsequent classification.

Image-Level Classification

The final classification is performed using an image-level XGBoost classifier[5], which takes the aggregated feature vector as input and produces a binary decision (real vs fake)[4].

Overall, the Geo-DefakeHop pipeline follows a hierarchical structure combining patch-based processing, multi-scale feature extraction, channel-wise selection, and feature aggregation. The resulting representation is used for image-level classification.

4.3.2. Test Results of the Geo-DefakeHop Model

This section presents the performance of the Geo-DefakeHop model on the selected datasets using both quantitative metrics and qualitative visualisations. The model is trained and evaluated separately on each dataset.

Evaluation metrics include accuracy, precision, recall, and F1-score. In addition, patch-level probability maps are generated from patch-wise prediction scores and arranged according to the original image layout.

Performance Evaluation on the FSI Dataset

The Geo-DefakeHop model is evaluated on the FSI dataset (CycleGAN-generated remote sensing images dataset).

Table 4.5.: Performance metrics of Geo-DefakeHop on the FSI dataset

Class	Precision	Recall	F1-score
Fake	0.94	0.97	0.96
Real	0.97	0.94	0.96
Accuracy		0.96	

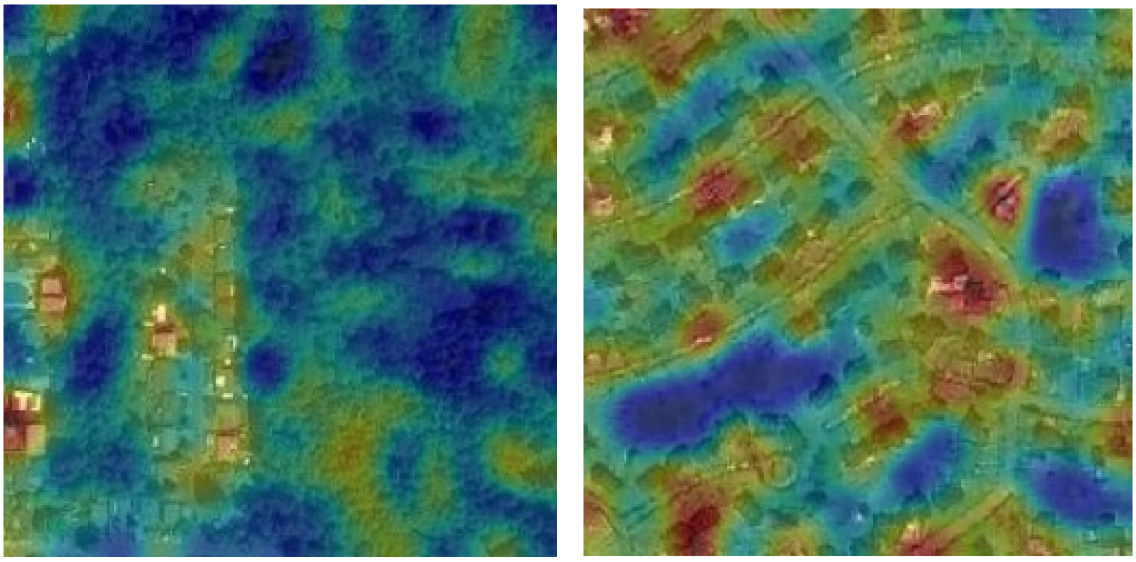
The obtained results are summarised in Table 4.5. The model achieves an overall accuracy of 96%, with comparable precision, recall, and F1-scores across both classes.

Table 4.6.: Confusion matrix of Geo-DefakeHop for the FSI dataset

	Predicted Fake	Predicted Real
Actual Fake	194	6
Actual Real	12	194

The confusion matrix in Table 4.6 shows that 6 fake images are misclassified as real and 12 real images are misclassified as fake. These results indicate that misclassifications occur in both classes, with slightly more errors in real samples than fake ones.

Representative patch-level probability maps are shown in Figure 4.4. In real samples, the responses appear relatively uniform, with high probabilities assigned across most re-



(a) GT: Real, Pred: Real, $P(\text{Real}) = 0.997$ (b) GT: Fake, Pred: Fake, $P(\text{Real}) = 0.185$

Figure 4.4.: Patch-level probability maps generated by the Geo-DefakeHop model for representative real and fake samples from the FSI dataset. Warmer regions indicate a lower probability of being real.

gions. In contrast, fake samples exhibit more localised variations, with regions of lower probability appearing as warmer areas in the heatmap.

Performance Evaluation on the FSI and DM-AER Datasets

The Geo-DefakeHop model is evaluated on a combined dataset consisting of FSI and DM-AER samples.

Table 4.7.: Performance metrics of Geo-DefakeHop on the FSI and DMAER dataset

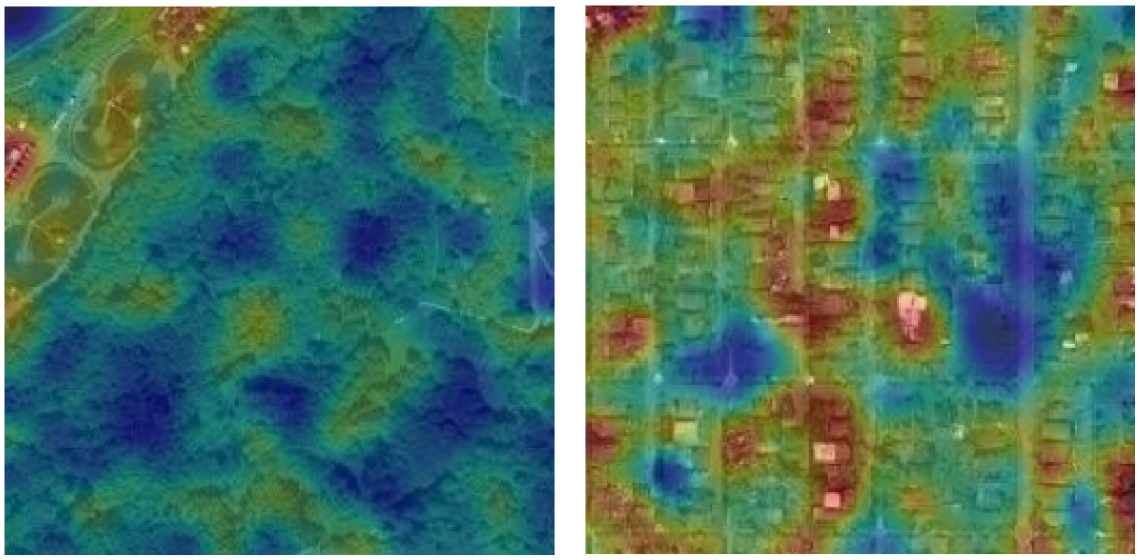
Class	**Precision**	**Recall**	**F1-score**
Fake	0.90	0.79	0.84
Real	0.81	0.92	0.86
Accuracy		0.85	

The results are summarised in Table 4.7. The model achieves an overall accuracy of 85%, which is lower than that obtained on the FSI dataset. The precision and recall values indicate differing behaviour across classes, with higher recall for real samples and lower recall for fake samples.

Table 4.8.: Confusion matrix of Geo-DefakeHop for the FSI and DMAER dataset

	Predicted Fake	**Predicted Real**
Actual Fake	201	55
Actual Real	22	240

The confusion matrix shown in Table 4.8 indicates that 55 fake samples were misclassified as real, while 22 real samples were misclassified as fake. Compared to the FSI dataset, a higher number of misclassifications is observed. In particular, fake samples are more frequently misclassified as real, indicating an asymmetry in error distribution.



(a) (GT: Real, Pred: Real

(b) GT: Fake, Pred: Fake

Figure 4.5.: Patch-level probability maps generated by the Geo-DefakeHop model for representative samples from the DM-AER dataset. Warmer regions indicate lower probability of being real.

Representative patch-level probability maps for the DM-AER dataset are shown in Figure 4.5. Real samples exhibit relatively uniform responses across the image, whereas fake samples show more localised variations, with regions of lower probability appearing as warmer areas.

Performance Evaluation on the CMF Dataset

The Geo-DefakeHop model is evaluated on the CMF dataset, which contains spatially localised manipulations.

Table 4.9.: Performance metrics of Geo-DefakeHop on the CMF dataset

Class	Precision	Recall	F1-score
Fake	0.83	0.73	0.78
Real	0.75	0.84	0.80
Accuracy		0.79	

The obtained results are summarised in Table 4.9. The model achieves an overall accuracy of 79%, which is lower than that observed for the previous datasets. The class-wise metrics show differing performance across classes, with higher recall for real samples than for fake ones. The lower recall for fake samples indicates that some manipulated images are not correctly identified.

Table 4.10.: Confusion matrix of Geo-DefakeHop for the CMF dataset

	Predicted Fake	Predicted Real
Actual Fake	63	23
Actual Real	13	70

The confusion matrix shown in Table 4.10 indicates that 23 fake samples are misclassified as real, while 13 real samples are misclassified as fake. This distribution shows that misclassifications occur more frequently for fake samples than for real ones.

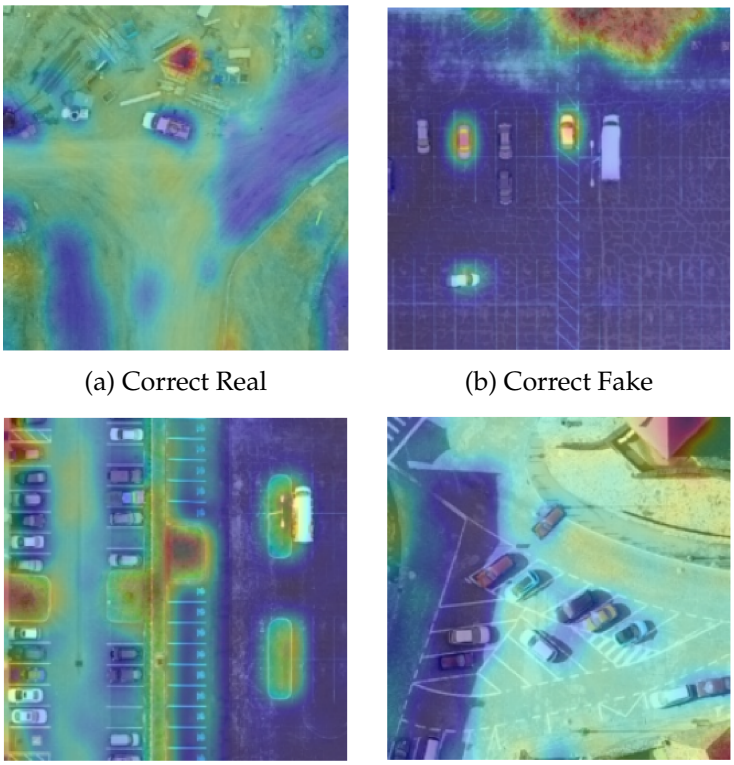
Figure 4.6 presents representative examples, including both correctly classified and misclassified samples. The activation patterns show that responses are often concentrated in regions with similar structures. In correctly classified fake samples, localised variations are more apparent. In contrast, misclassified cases exhibit more ambiguous patterns, in which similar responses may also appear in real images, whereas some manipulated regions do not produce clearly distinguishable responses.

4.4. Discussion

The results presented in this chapter indicate that detection performance varies with the choice of feature representation and its relationship to the underlying manipulation characteristics[35].

For GAN-generated images, the FFT-based ResNet achieves high performance, consistent with prior observations that such data contain global frequency-domain artefacts[41, 12]. This suggests that frequency-based representations can capture these global inconsistencies, which may be less apparent in the spatial domain.

In contrast, performance on the CMF dataset is relatively lower, particularly for the wavelet-based ResNet. Although the wavelet representation preserves spatial locality and supports the analysis of localised patterns[24, 20], the results indicate that these patterns are not always sufficiently distinctive for reliable detection. This is also reflected in the



(a) Correct Real (b) Correct Fake

(c) Real → Fake (d) Fake → Real

Figure 4.6.: Patch-level probability maps for representative samples from the CMF dataset, including both correctly classified and misclassified cases. Warmer regions indicate a lower probability of being real.

confusion matrix and Grad-CAM results, where natural repetitive structures in real images yield responses similar to manipulated regions[8, 10].

The Geo-DefakeHop model provides an alternative perspective through explicit analysis of frequency-selective components[4]. Its performance remains relatively strong on GAN-generated datasets but decreases on mixed and CMF data, suggesting that its effectiveness depends on the presence of consistent and discriminative frequency patterns[7].

Overall, the results show that different representations exhibit varying levels of effectiveness depending on the manipulation type. Representations that emphasise global frequency information perform well for GAN-generated artefacts, whereas localised representations are more relevant for spatial manipulations, although with increased ambiguity[35, 11].

These observations also point to a trade-off between performance and interpretability. Deep learning models such as ResNet achieve strong results when aligned with suitable representations but provide limited insight into the decision process[41]. In contrast,

Geo-DefakeHop enables interpretable, channel-wise analysis, while its performance varies across datasets[4]. These findings suggest that relying on a single representation may be insufficient for handling diverse manipulation types and that combining complementary representations could be a promising approach to improve robustness.

The next chapter further investigates the generalisation behaviour of the proposed models across datasets, focusing on cross-dataset performance and representation-specific limitations.

Part IV.

Experimental Analysis and Results

5. Analytical Tools and Evaluation Metrics

To better understand the behaviour of the proposed models beyond standard performance metrics, a set of interpretability-driven analyses is conducted. Rather than relying on accuracy-based evaluation alone, this section investigates two key aspects: how different models respond to manipulation characteristics, and which feature representations they exploit during decision-making.

5.1. Interpretability and Analysis Framework

To achieve this, two complementary analysis strategies are employed, tailored to the structure of each model.

5.1.1. ResNet-Based Interpretability Analysis

For the ResNet-based models[41], a gradient-based interpretability approach is used to identify the features driving the model's predictions. In particular, Grad-CAM generates activation maps by backpropagating gradients of the target class through the final convolutional layer[31].

Unlike conventional spatial-domain applications, the analysis is performed on transformed representations, namely FFT- and wavelet-based inputs. To interpret these representations, a feature-specific backprojection strategy is applied.

For the FFT-based model, Grad-CAM maps are computed in the frequency domain. They are then projected back to the spatial domain via an inverse Fourier transform, enabling localisation of spectral artefacts. The frequency spectrum is also divided into low-, mid-, and high-frequency bands. Their contributions are quantified based on Grad-CAM responses.

For the wavelet-based model, Grad-CAM is applied to multi-channel wavelet representations. Activation maps are then reconstructed in the spatial domain using inverse discrete wavelet transformation, enabling analysis of localised sub-band responses.

This combined approach provides both quantitative and visual insights, linking spectral components to manipulation artefacts in the image domain.

5.1.2. Geo-DefakeHop Feature Analysis

For the Geo-DefakeHop model, an interpretable analysis framework is employed to examine the contribution of frequency-selective components. Unlike gradient-based ap-

proaches, this model allows direct inspection of learned representations through its channel-wise and multi-hop structure[4].

Channel-wise discriminative analysis is performed using metrics such as the F1-score. This helps to identify informative frequency components. In addition, channel response distributions are analysed to evaluate the separability between real and fake samples.

Multi-hop feature behaviour is examined through visualisation of Saab filter responses, revealing how different representation levels capture distinct structural patterns. Furthermore, patch-level probability maps are used to analyse the spatial distribution of model predictions by aggregating patch-wise outputs.

Together, these analyses provide an interpretable understanding of model behaviour and highlight strengths and limitations across different manipulation types. Additionally, cross-dataset and cross-manipulation evaluations are conducted to assess generalisation behaviour. These evaluations determine whether learned representations are transferable or manipulation-specific.

5.2. Analysis of Learned Representations in ResNet for GAN-generated Fake Images

Although the FFT-based ResNet model achieves high classification performance, it is essential to analyse whether this performance is driven by meaningful manipulation-related features or by unintended biases. To this end, Grad-CAM-based spectral analysis[31] is employed to identify the frequency components that contribute to the model's decisions.

5.2.1. Frequency Band Contribution Analysis

To quantify how the model utilises frequency information, the Fourier spectrum is divided into three bands: low, mid, and high frequency[41]. The contribution of each band is computed using Grad-CAM activation maps by averaging the activations within its corresponding frequency region. These values reflect the relative importance assigned by the model to each frequency band during decision-making, rather than representing spectral energy[31].

Table 5.1.: Average frequency band contributions for real and fake images based on Grad-CAM analysis

Class	Low	Mid	High
Real	0.020	0.017	0.031
Fake	0.242	0.333	0.294

As shown in Table 5.1 and Figure 5.1, fake images exhibit significantly stronger activations in the mid- and high-frequency bands compared to real images. In particular, the

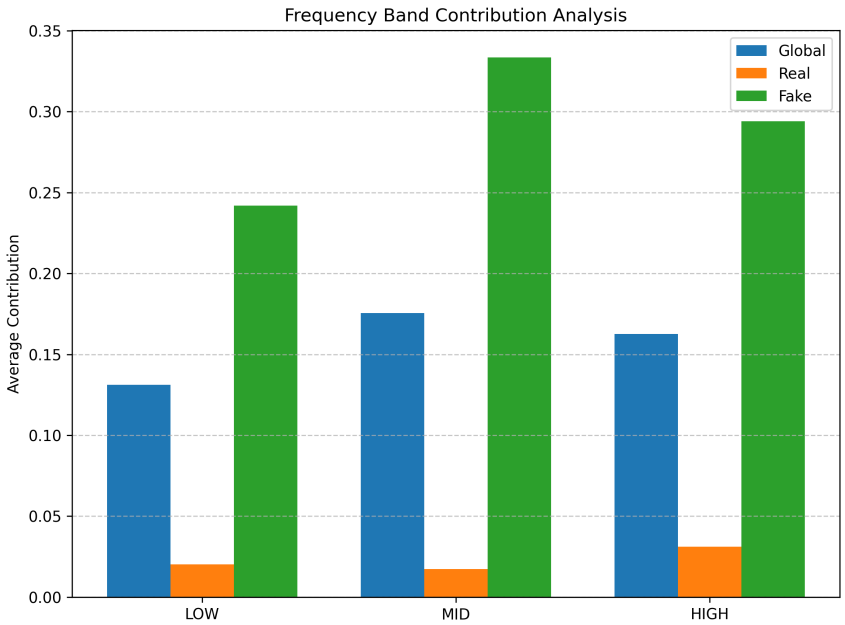


Figure 5.1.: Average frequency band contribution based on Grad-CAM activations for real and fake images.

mid-frequency band shows the highest contribution for fake samples (0.333), followed by the high-frequency band (0.294), whereas real images remain consistently low across all frequency ranges.

This pattern indicates that the model focuses on structured spectral irregularities rather than uniformly distributed noise. In contrast, real images do not exhibit consistent frequency-domain patterns, leading to weaker, more diffuse activations[12][41].

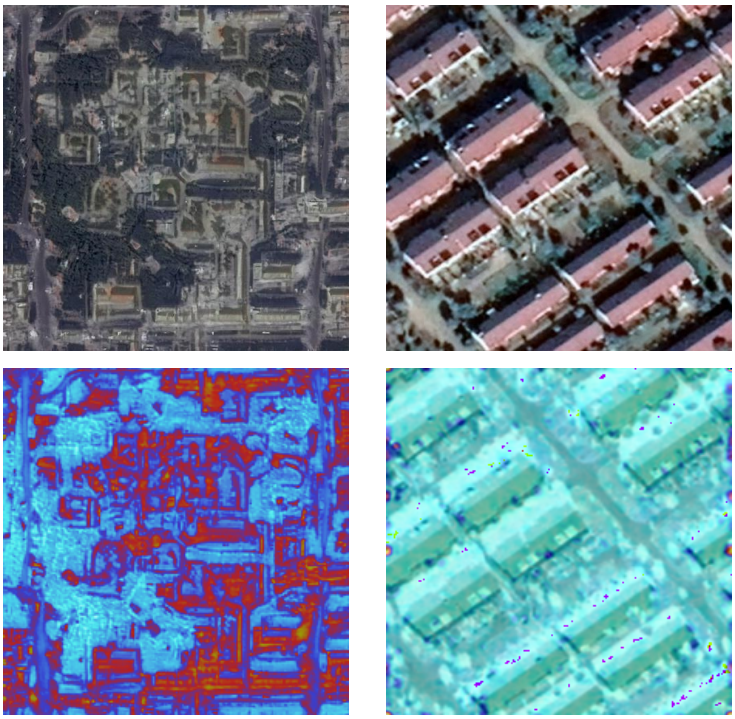
Overall, the results demonstrate that the effectiveness of the FFT-based representation stems from its ability to expose manipulation-specific spectral inconsistencies[35].

5.2.2. Sample-Level Analysis

Grad-CAM visualisations for representative real and fake images are presented in Figure 5.2.

The fake image exhibits strong and spatially structured activations, concentrated around edges and fine details. These regions align with dominant mid- and high-frequency contributions. In contrast, the real image produces weak, diffuse activations with no consistent regions of emphasis.

This difference indicates that the model does not respond to isolated artefacts, but rather to structured patterns distributed across the image. The activations in fake samples appear globally organised, whereas real images lack such a consistent structure[41].



(a) Fake image (CycleGAN-generated) with strong high-frequency activation. Band contributions: Low = 0.13, Mid = 0.30, High = 0.30

(b) Real image with weak and unstructured responses. Band contributions: Low = 0.00, Mid = 0.00, High = 0.05

Figure 5.2.: Grad-CAM visualisations for representative real and fake images.

Discussion

The results indicate that the model's performance is driven by structured frequency-domain artefacts rather than random patterns, confirming the effectiveness of FFT-based representations for GAN-generated images[41, 12].

However, this also reveals a key limitation: the model relies on the presence of such spectral inconsistencies, which may vary across different generative processes. As a result, its performance is inherently representation-dependent.

5.3. Analysis of Learned Representations in Geo-DefakeHop for GAN-generated Fake Images

To analyse how Geo-DefakeHop[4] distinguishes real and fake images, feature representations across different hops are examined. The model employs a multi-stage Saab transform, where each hop captures progressively more spectral and spatial characteristics[7].

Building on this, in the first stage (Hop A), the filters primarily extract low-level features such as edges and basic spatial structures, corresponding to low-frequency components[4]. In the second stage (Hop B), the representations become more structured and capture mid-frequency components, where local inconsistencies and texture irregularities introduced by generative models begin to emerge. In the final stage (Hop C), the features become more abstract and global, representing higher-level information while losing sensitivity to fine-grained details.

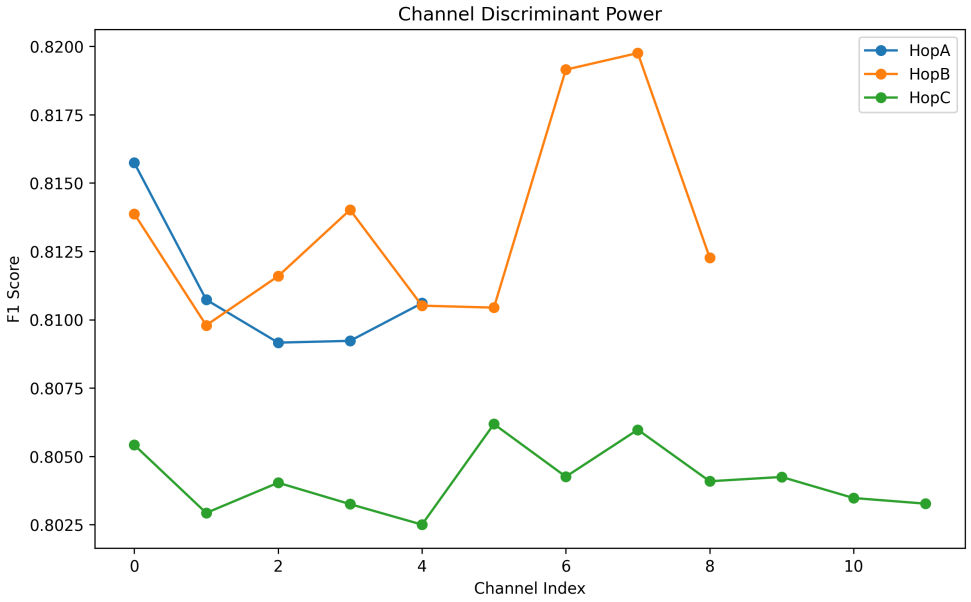
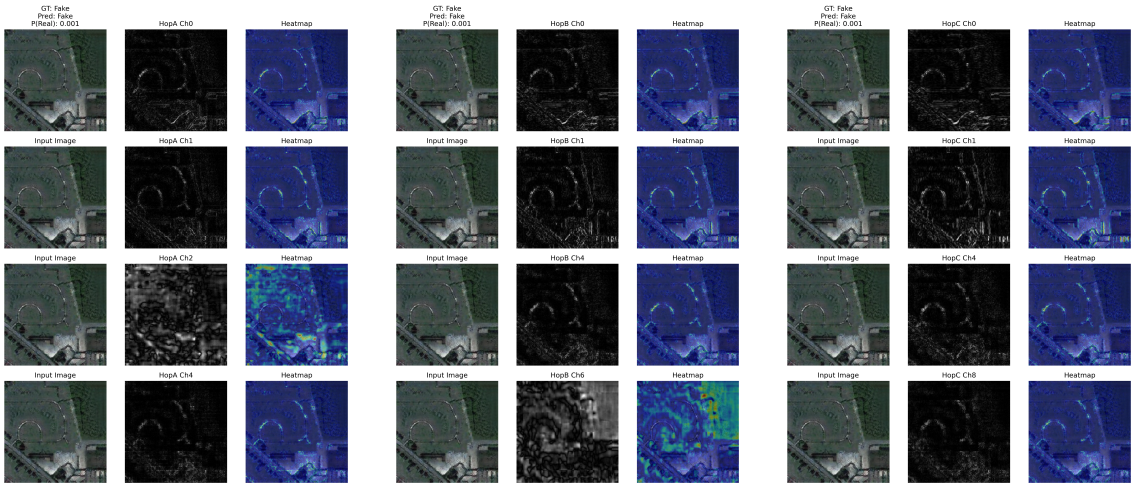


Figure 5.3.: Channel-wise discriminative performance across different hops in Geo-DefakeHop, measured using F1-scores for individual channels.

The discriminative capability of these representations is evaluated through channel-wise analysis. As shown in Figure 5.3, channels in Hop B achieve the highest F1-scores, followed by Hop A, while Hop C exhibits noticeably lower performance. This indicates that discriminative information is concentrated in intermediate frequency representations.

This suggests that mid-frequency components capture the most informative artefact patterns, whereas higher-level representations lose sensitivity to fine-grained inconsistencies. Unlike ResNet, Geo-DefakeHop explicitly decomposes these representations into inter-

pretable frequency-selective channels, enabling direct identification of their contribution[4].



(a) Hop A (b) Hop B (c) Hop C

Figure 5.4.: Visualization of Saab filter responses and activation maps across different hops of the Geo-DefakeHop model for a representative fake image.

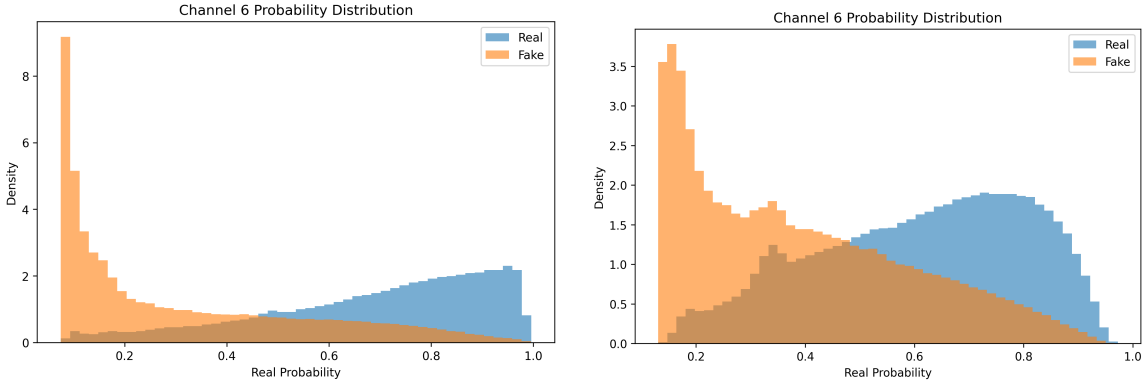
This behaviour is further supported by the visualisations in Figure 5.4. Hop B exhibits more localised and structured activations, particularly around textured regions and object boundaries, where generative artefacts are more likely to appear.

In contrast, Hop A primarily captures basic spatial structures, while Hop C produces more diffuse responses with reduced sensitivity to fine-grained details. This confirms that intermediate representations provide the most discriminative features for fake image detection.

CycleGAN and StyleGAN2-generated Images Differences in Geo-DefakeHop

The performance of Geo-DefakeHop decreases noticeably when evaluated on mixed datasets containing both CycleGAN-[45] and StyleGAN2[22]-generated fake images. As shown in Table 4.5 and Table 4.7, the model achieves high performance on the FSI dataset, with an accuracy of approximately 0.96. However, when evaluated on the combined FSI and DMAER datasets, the overall accuracy drops to 0.85, indicating reduced generalisation capability.

This degradation can be attributed to the fundamental differences between CycleGAN and StyleGAN2-generated fake images. CycleGAN, based on image-to-image translation, often introduces structural inconsistencies and blurred textures due to imperfect domain mapping[45]. In contrast, StyleGAN2 produces visually more realistic images with improved low-frequency fidelity, while embedding more subtle artefacts in specific frequency bands[17].



(a) for CycleGAN-generated dataset (FSI dataset)

(b) for StyleGAN2-generated dataset (DMAER dataset)

Figure 5.5.: Channel-wise probability distributions of Hop B Channel 6 for (a) CycleGAN-generated (FSI) and (b) StyleGAN2-generated (DM-AER) datasets.

To investigate this behaviour, the discriminative power of individual channels is analysed using channel-wise probability distributions. As shown in Figure 5.5, Hop B Channel 6 clearly separates real and fake samples in the FSI dataset, indicating strong discriminative capability. In contrast, for the DMAER dataset, the distributions exhibit significant overlap, indicating reduced separability.

This suggests that frequency-based features are highly effective for detecting artifacts introduced by CycleGAN, while their discriminative power decreases for more advanced generative models such as StyleGAN2, where synthetic images exhibit more realistic and less distinguishable frequency characteristics[17].

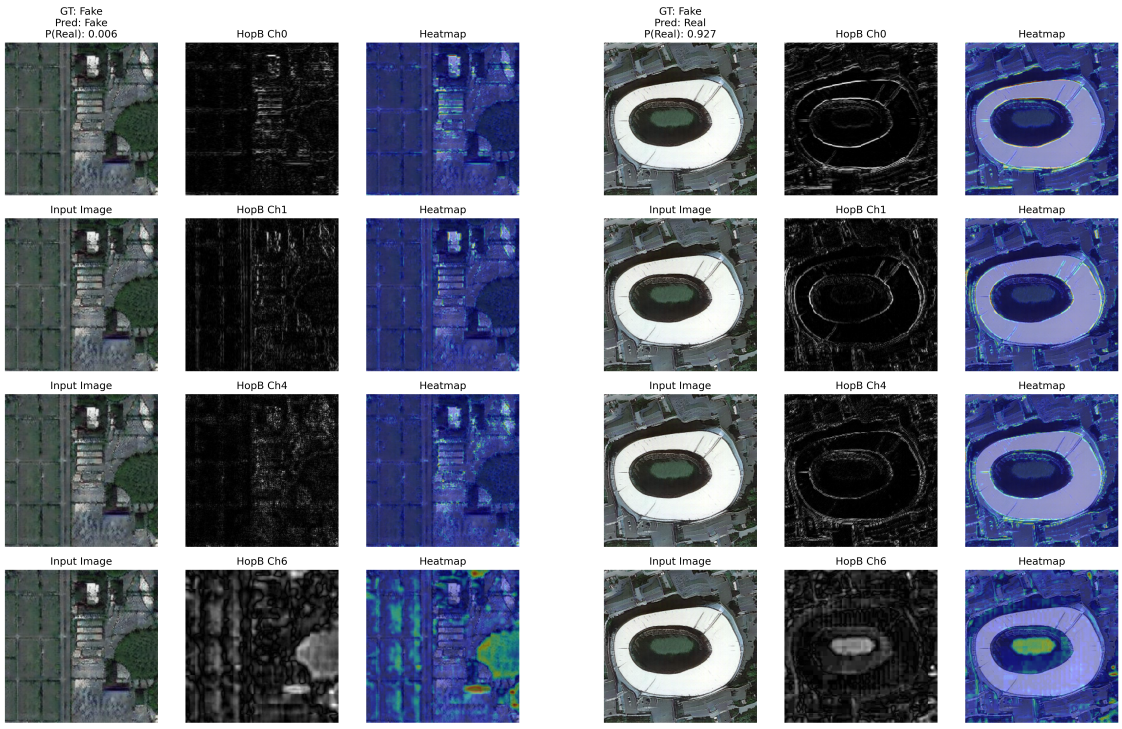
This behaviour is further supported by the Saab feature visualisations in Figure 5.6. For CycleGAN-generated images, the responses are strong and structured, clearly highlighting artefact-related inconsistencies. In contrast, StyleGAN2-generated images produce weaker and more diffuse activations, making artefact localisation more challenging.

This difference indicates that StyleGAN2 produces more realistic spectral characteristics, thereby reducing the discriminative power of frequency-based features and potentially leading to misclassification[17].

These observations indicate that Geo-DefakeHop relies on the presence of structured and separable frequency patterns, particularly in intermediate representations[4]. When such patterns are less pronounced, as in StyleGAN2, the model's discriminative capability decreases significantly.

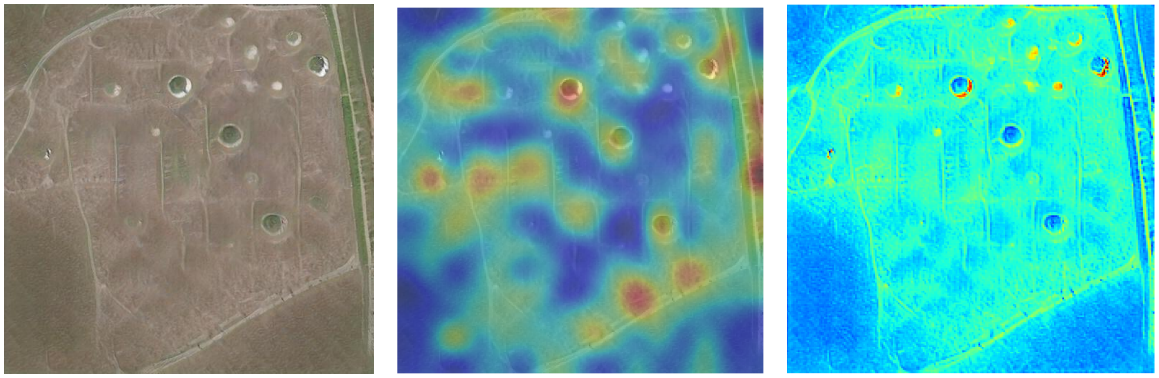
Figure 5.7 further highlights this limitation. While Geo-DefakeHop produces weak, diffuse activations that lead to misclassification, ResNet successfully detects the image by capturing subtle, distributed frequency inconsistencies.

This contrast demonstrates a fundamental difference between the two models. Geo-



(a) CycleGAN-generated fake image (b) StyleGAN2-generated fake image

Figure 5.6.: Comparison of Saab filter responses in Hop B for fake images generated by (a) CycleGAN and (b) StyleGAN2.



(a) StyleGAN-generated fake image (DM-AER dataset) (b) Geo-DefakeHop Heatmap (Pred: Real) (c) FFT-ResNet Heatmap(Pred: Fake)

Figure 5.7.: Comparison of Geo-DefakeHop and FFT-ResNet heatmaps for a StyleGAN2-generated image.

DefakeHop relies on explicit, separable frequency structures[4], whereas ResNet learns more flexible, distributed representations, enabling stronger generalisation across different generative models[41].

This difference can also be explained by the underlying representation mechanisms of the two models. Geo-DefakeHop relies on Saab filters, which are data-driven but fixed once extracted[4]. As a result, the model operates on a predefined set of linear subspace decompositions and may struggle to adapt to subtle and complex statistical variations introduced by more advanced generative models such as StyleGAN2.

In contrast, ResNet is trained end-to-end, allowing it to learn hierarchical, nonlinear feature representations directly from the data. This flexibility enables the model to capture more complex and distributed statistical patterns that may not be explicitly represented in fixed transform-based decompositions. Consequently, ResNet demonstrates stronger generalisation when dealing with more realistic, less structured artefacts.

Overall, these findings confirm that the effectiveness of a detection model is strongly dependent on the compatibility between its feature representation and the characteristics of the underlying manipulation. This observation further motivates the use of complementary approaches to achieve robust and generalisable fake image detection.

Discussion

The results highlight a fundamental difference in how Geo-DefakeHop and ResNet capture manipulation artefacts for GAN-generated fake remote sensing images. Geo-DefakeHop relies on explicitly separable, frequency-selective channels—particularly in intermediate representations—providing clear interpretability and direct insight into which features drive detection[4]. In contrast, FFT-based ResNet learns more distributed and flexible representations, enabling stronger generalisation across different generative models[41].

However, this interpretability comes with a limitation. Geo-DefakeHop relies on the presence of structured, distinguishable frequency artefacts, which are prominent in CycleGAN but significantly weaker in more advanced models such as StyleGAN2. As a result, its discriminative capability decreases when such patterns become less separable.

The channel-wise structure of Geo-DefakeHop enables direct comparison of feature responses across different frequency bands. This allows systematic analysis of a wide range of artefact types. It is not limited to high-frequency inconsistencies, but also covers mid- and low-frequency patterns. Such capability is especially valuable for future investigations. Different generative models may introduce artefacts in diverse spectral regions.

These findings indicate that detection performance is not solely model-dependent, but is also strongly influenced by the compatibility between the feature representation and the underlying manipulation characteristics.

Overall, these results confirm the central hypothesis of this work. Detection performance is affected by the chosen feature representation. They further motivate the use of complementary approaches. Combining the robustness of deep learning models with the

interpretability of transform-based methods achieves more reliable and generalisable fake image detection.

5.4. Cross-Dataset Generalisation Analysis

To evaluate the generalisation capability of the proposed models, cross-dataset experiments are conducted by training models on GAN-generated images and evaluating them on a different manipulation type, namely copy-move forgery (CMF). This setting enables the assessment of whether representations learned from GAN data can generalise to fundamentally different manipulation characteristics.

Table 5.2.: Performance of GAN-generated images trained models evaluated on CMF dataset

Model	Accuracy	Precision	Recall	F1 Score
Geo-DefakeHop	0.500	0.500	0.920	0.648
FFT-based ResNet	0.465	0.479	0.810	0.602

(a) Original CMF image, GT: Fake

(b) FFT-based ResNet heatmap, Pred: Real

(c) Geo-DefakeHop heatmap (GAN-trained), Pred: Real

Figure 5.8.: Model responses on a CMF image. Both models are trained on GAN-generated data and fail to capture CMF-specific patterns, leading to incorrect predictions.

As shown in Table 5.2, both models exhibit a significant performance degradation when evaluated on CMF images. While the recall values remain relatively high (0.92 for Geo-DefakeHop and 0.81 for FFT-based ResNet), the overall accuracy and precision are considerably low.

This distinction indicates that the models tend to favour the *real* class, leading to many fake samples being misclassified. As a result, the models achieve high recall primarily by correctly identifying real samples rather than by effectively detecting manipulated images.

This behaviour reflects a fundamental mismatch between the learned representations and the characteristics of CMF manipulations. Models trained on GAN-generated images rely heavily on frequency-domain artefacts, which are largely absent in CMF data. In contrast, CMF forgeries are characterised by spatial duplication and structural repetition, requiring the detection of intra-image similarities rather than spectral inconsistencies[44].

This mismatch is further illustrated in Fig. 5.8. Both models focus on high-frequency regions and textured structures, while failing to capture the duplicated areas that define CMF manipulation. Since most of the image remains unchanged and appears natural, the models interpret it as real. Consequently, both models produce non-discriminative activations and misclassify the image as real.

Due to this representational mismatch, frequency-based models is not sufficient to capture the relevant cues necessary for CMF detection. As a result, their predictions become biased and unreliable when applied to this manipulation type.

These findings demonstrate that models trained on GAN-generated images are insufficient for detecting fundamentally different types of manipulation. This limitation highlights the need for alternative representations and motivates the development of specialised models for CMF detection, as explored in the following section.

5.5. Analysis of Learned Representations on Copy-Move Forgery (CMF)

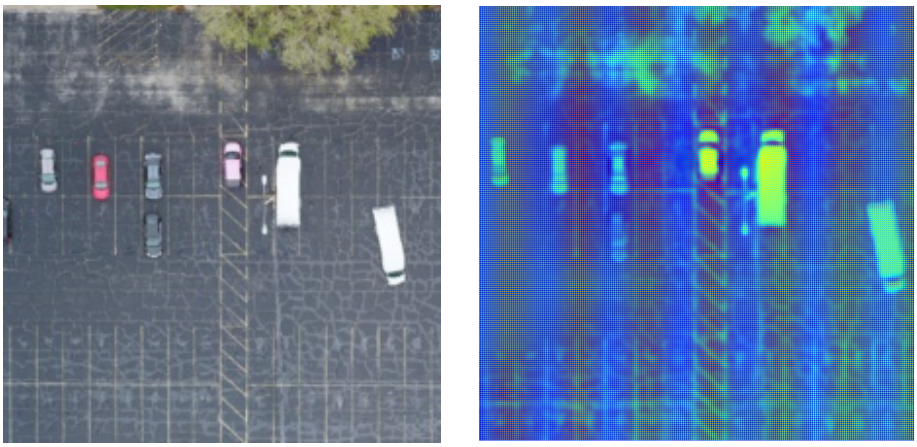
Following the cross-dataset analysis in Section 5.4, it is evident that models trained on GAN-generated images show reduced generalisation to CMF data due to a representation mismatch. In this section, we further investigate how different models behave when explicitly trained and evaluated on CMF images, with the aim of understanding the characteristics required for effective CMF detection.

5.5.1. Wavelet and ResNet-based Analysis

Copy-move forgery (CMF) differs fundamentally from GAN-generated images in terms of their characteristic features. Instead of introducing synthetic content, CMF duplicates regions within the same image, preserving overall spectral consistency while altering spatial structure[8]. As a result, CMF does not introduce strong global frequency artefacts, making it inherently more challenging to detect using frequency-domain analysis alone[10].

This distinction makes frequency-domain representations insufficient for CMF detection. Unlike GAN-generated images, CMF images do not exhibit strong spectral inconsistencies, limiting the effectiveness of FFT-based approaches[10].

Instead, effective CMF detection requires the model to identify intra-image similarities, as the manipulation is defined by duplicated regions within the same image[44]. This shifts the detection problem from identifying global anomalies to recognising structurally similar patterns across different spatial locations.



(a) Input Image (CMF) (b) ResNet Heatmap

Figure 5.9.: Grad-CAM visualisation of ResNet for a CMF image. (a) Input image. (b) Corresponding activation map highlighting duplicated regions.

To address this limitation, wavelet-based representations are employed. Unlike Fourier transforms, wavelets provide joint spatial-frequency localisation, enabling the detection of local inconsistencies such as duplicated structures and boundary artefacts[26].

This behaviour is clearly observed in the ResNet activation maps (Figure 5.9). The wavelet-based ResNet model focuses on localised regions corresponding to structurally similar areas within the image. These activation patterns indicate that the model relies on intra-image similarity rather than global artefacts.

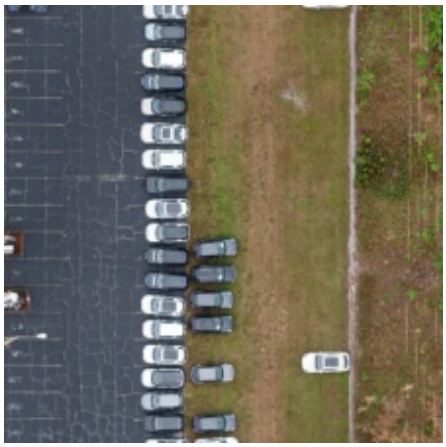
This similarity-based detection strategy introduces an important limitation. As illustrated in Figure 5.10, natural repetitive patterns in real images can produce similar responses, leading to false positive predictions. This suggests that spatial similarity alone is not a reliable indicator of manipulation[44].

Overall, while wavelet-based representations improve CMF detection, they introduce ambiguity due to the inherent repetition in natural scenes.

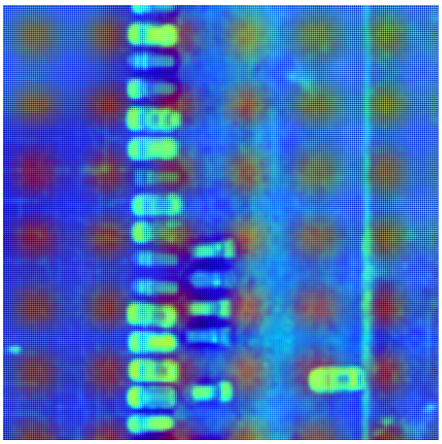
5.5.2. Geo-DefakeHop on CMF Images

Unlike the ResNet-based approach, Geo-DefakeHop is applied to CMF images without architectural modifications. While Section 4 provides its quantitative performance, here we analyse its behaviour through feature representations and channel-wise responses.

Geo-DefakeHop captures frequency-selective characteristics through channel-wise statistical analysis[4]. This is effective for GAN-generated images that exhibit structured spectral artefacts. However, for CMF data, which lacks such frequency inconsistencies, the detection task shifts toward identifying intra-image similarity rather than spectral differences[44].



(a) Input Image (Real)



(b) ResNet Heatmap (Pred: Fake)

Figure 5.10.: Grad-CAM visualisation of a misclassified real image. (a) Input image. (b) Activation map highlighting repetitive structures.

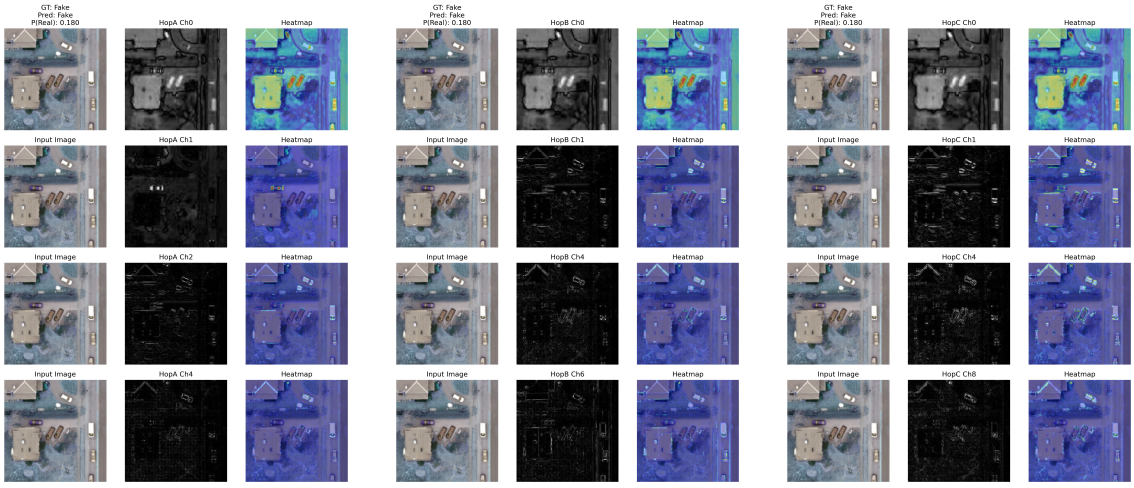
This behaviour is reflected in Fig. 5.11. Hop A captures low-level spatial structures, emphasising edges, object boundaries, and visually similar regions. These responses align well with the duplicated patterns characteristic of CMF. In contrast, Hop B produces more texture-oriented representations[7], where similarity patterns become less explicit, while Hop C yields more abstract and diffuse responses with limited spatial correspondence. This indicates that early-stage representations, particularly Hop A, are the most relevant for capturing similarity-based cues in CMF detection.

Channel-wise analysis in Fig. 5.13 further supports this observation, showing that Hop A channels consistently achieve higher F1 scores than those in later stages. This highlights the importance of low-level spatial features for CMF detection.

However, this behaviour introduces a key limitation. As shown in Fig. 5.12, the model misclassifies real images with natural repetitive structures as fake. The activation maps indicate strong responses to visually similar regions, leading the model to interpret natural repetition as manipulation.

In addition, Fig. 5.14 illustrates a false negative case, where similarity cues are weak or ambiguous. In such cases, the model fails to distinguish manipulated regions from the background. These results reveal a fundamental limitation: while Geo-DefakeHop is sensitive to spatial similarity, it lacks selectivity. It tends to overreact to natural repetition while missing subtle manipulations.

Overall, this analysis demonstrates a clear behavioural shift. Unlike GAN detection, which relies on frequency inconsistencies, CMF detection requires capturing intra-image similarity. Although Geo-DefakeHop partially adapts via early-stage features[4], its frequency-oriented design limits its robustness against copy-move forgery.



(a) Hop A (Low-level features) (b) Hop B (Mid-level features) (c) Hop C (High-level features)

Figure 5.11.: Visualisation of Saab filter responses across Hop A, Hop B, and Hop C of the Geo-DefakeHop model on a CMF image. Selected channels and their activation maps are shown for each hop.

5.5.3. Discussion and Motivation for Ensemble Learning

The analyses in this section demonstrate that different types of fake images exhibit fundamentally distinct characteristics, which directly affect the behaviour and effectiveness of detection models[35][8][41].

For GAN-generated images, both ResNet and Geo-DefakeHop exploit frequency-based inconsistencies, but in different ways. ResNet learns robust, distributed frequency-aware representations, enabling strong generalisation across different generative models[41]. In contrast, Geo-DefakeHop provides explicit interpretability by decomposing frequency components and capturing discriminative mid-frequency patterns[4]. However, its effectiveness depends on the presence of structured spectral artefacts, making it sensitive to variations across models such as CycleGAN and StyleGAN2.

On the other hand, copy-move forgery (CMF) does not introduce global spectral inconsistencies, rendering frequency-based representations less effective; thus, detection instead relies on spatial-similarity cues[44]. ResNet adapts by focusing on structurally similar regions but may produce false positives in the presence of natural repetition. Geo-DefakeHop captures such similarities through low-level representations (Hop A), but lacks robustness and selectivity, leading to both false positives and false negatives.

Collectively, these findings highlight a fundamental limitation: each model is inherently biased toward a specific feature representation[13]. Frequency-based approaches are effective when spectral artefacts are present, while spatially sensitive representations are better suited for similarity-based manipulations. However, no single representation is sufficient

(a) Example 1: A real image misclassified as fake (P(Real)=0.341).

(b) Example 2: Another real image misclassified as fake (P(Real)=0.495).

Figure 5.12.: False-positive cases of Geo-DefakeHop on CMF images. Hop A Channel 0 responses are visualised.

for robust detection across all types of fake images.

As a result, there is a need for a unified detection framework that leverages complementary strengths. By combining models that capture both frequency inconsistencies and spatial similarities, it becomes possible to improve robustness and generalisation[28].

Based on this insight, an ensemble-based approach is proposed that integrates multiple models to exploit their complementary behaviours. By aggregating their predictions, the ensemble mitigates individual limitations and enables more reliable detection across diverse manipulation types, including GAN-generated and copy-move forged images.

Therefore, fake image detection should not be treated as a single-domain classification task but rather as a multi-characteristic problem that requires adaptive, complementary feature representations.

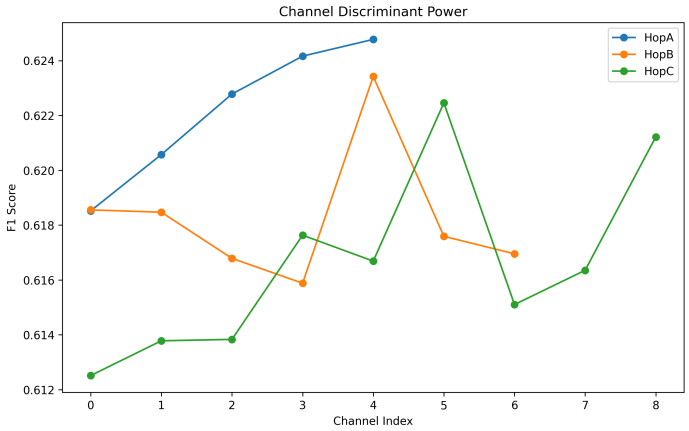


Figure 5.13.: Channel-wise F1 scores across different hops of Geo-DefakeHop on CMF data.

5.5.4. Cross-Manipulation Evaluation: CMF to GAN

To further analyse generalisation behaviour, an additional cross-manipulation experiment is conducted by evaluating CMF-trained models on GAN-generated images. This setting complements the previous analysis and enables a bidirectional assessment of representation transferability.

As shown in Table 5.3, both Geo-DefakeHop (CMF) and the wavelet-based ResNet exhibit a significant performance degradation when applied to GAN-generated data. In particular, Geo-DefakeHop achieves an accuracy of 0.44, with an extremely low F1-score of 0.08 for the fake class.

This behaviour is further illustrated in the confusion matrices (Table 5.4). Geo-DefakeHop correctly identifies only 4 out of 85 fake samples, misclassifying the majority as real. This indicates a strong bias toward the real class when the expected spatial similarity cues are absent.

Table 5.3.: Cross-manipulation evaluation: Performance of CMF-trained models on GAN-generated images.

Model	**Accuracy**	**F1 (Fake)**	**F1 (Real)**	**TPR**	**TNR**
Geo-DefakeHop (CMF)	0.44	0.08	0.60	0.05	0.84
ResNet (Wavelet, CMF)	0.54	0.48	0.59	0.41	0.67

In contrast, when evaluated on data aligned with their training distribution, these models exhibit more stable and balanced performance. This confirms that model behaviour is strongly influenced by the characteristics of the training data.

More importantly, this experiment reveals that the generalisation failure is not unidi-

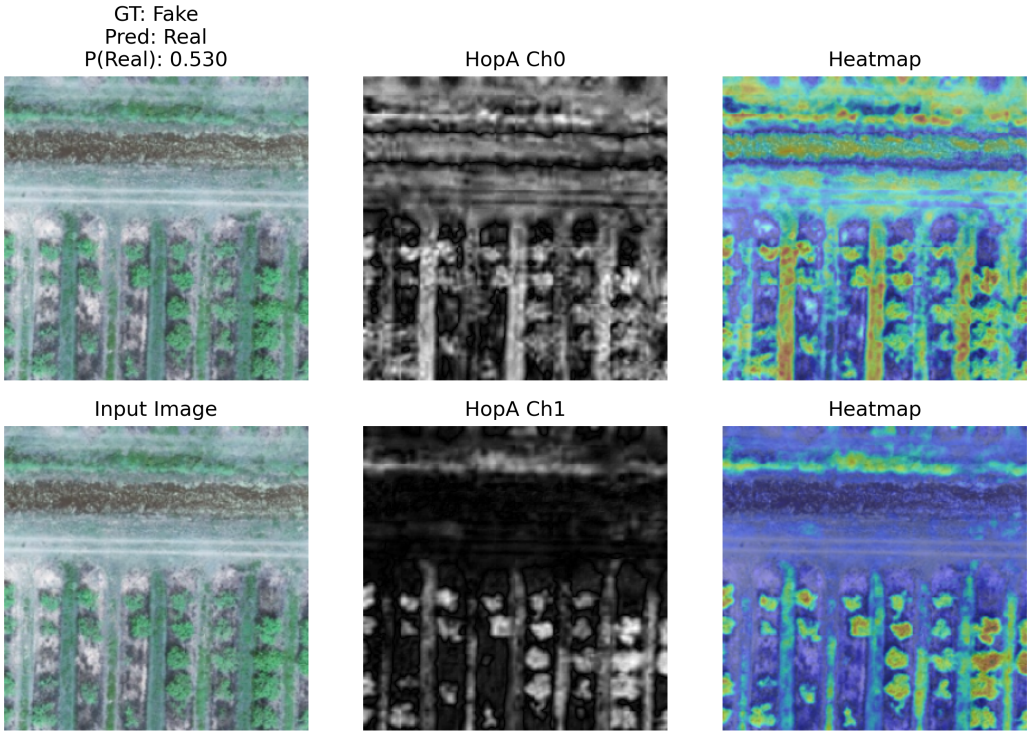


Figure 5.14.: False negative case where the model fails to detect manipulation.

rectional. While GAN-trained models is not sufficient on CMF data (Section 5.4), CMF-trained models also do not have enough performance on GAN-generated images. This indicates that the learned representations are inherently manipulation-specific rather than generalisable.

These findings provide strong evidence that different manipulation types exhibit distinct and partially non-overlapping characteristics. GAN-generated images are characterised by frequency-domain artefacts, whereas CMF images rely on spatial duplication and structural similarity.

As a result, models trained to capture one type of artefact develop a representational

Table 5.4.: Confusion matrices showing the prediction behaviour of CMF-trained models on GAN-generated images.

	Geo-DefakeHop (CMF)		**ResNet (Wavelet, CMF)**	
	Fake	**Real**	**Fake**	**Real**
Fake (GT)	4	81	35	50
Real (GT)	13	70	27	56

bias, which limits their applicability to other manipulation types. This further supports the need for a representation-aware approach that can leverage complementary models for robust detection.

5.6. Discussion: Representation-Dependent Generalisation

The results presented throughout this section reveal a fundamental limitation of fake remote sensing image detection: *models have difficulties learning universal forgery representations, but rather manipulation-specific characteristics*.

Table 5.5.: Cross-dataset evaluation of the proposed models. Models are trained on one dataset and evaluated on a different manipulation type.

Model	Training Dataset	Testing Dataset	Accuracy
FFT-ResNet	GAN	GAN	0.98
FFT-ResNet	GAN	CMF	0.46
Wavelet-ResNet	CMF	CMF	0.68
Wavelet-ResNet	CMF	GAN	0.54
Geo-DefakeHop	GAN	GAN	0.96
Geo-DefakeHop	GAN	CMF	0.5
Geo-DefakeHop (CMF)	CMF	CMF	0.79
Geo-DefakeHop (CMF)	CMF	GAN	0.44

This behaviour is most evident in Table 5.5. All models achieve relatively high accuracy on the same manipulation type they are trained on (e.g., FFT-ResNet reaches 0.98 on GAN data). However, their performance drops sharply in cross-manipulation cases (e.g., dropping to 0.46 on CMF). This consistent pattern for all models shows that generalisation failure is not unique to any model but is rooted in the representation itself.

This limitation stems from differences between manipulation types. GAN-generated images introduce global frequency artefacts[41]. Copy-move forgeries show local spatial duplication[44]. As a result, models trained on frequency-based data miss similarity-driven patterns. Meanwhile, models focused on spatial structures ignore spectral inconsistencies.

Importantly, this limitation is not due to insufficient model capacity, but rather due to a *representation mismatch*. Each model learns features aligned with the characteristics of its training data, leading to strong performance within its domain but poor transferability beyond it. This confirms that fake image detection is inherently a *characteristic-dependent problem*[35].

Furthermore, the results suggest that not all manipulation types show the same level of detection difficulty. In particular, GAN-generated images are comparatively easier to detect. This is due to the presence of consistent and globally distributed frequency artefacts. In contrast, CMF images are more challenging[8]. They preserve overall image statistics

and introduce only subtle, localised structural inconsistencies[44]. This difference further reinforces the need for representation-specific approaches.

These findings challenge the common assumption that a single, unified model can robustly detect all types of fake images. Instead, the results show that different manipulation types need specialised detection strategies. Each one should be tailored to a specific feature domain.

At the same time, the observed failure patterns are not random. They are structured and complementary. Frequency-based models detect GAN artefacts, but are not sufficient on CMF. Spatially sensitive models show the opposite behaviour. This complementarity suggests that these models capture different, non-overlapping aspects of manipulation.

Therefore, rather than seeking a universal representation, a more effective approach is to integrate multiple specialised models. By combining frequency-aware and spatially-aware representations, it becomes possible to leverage their strengths. This integration helps overcome individual limitations[28][3].

In summary, the analysis in this section leads to three key insights:

- Fake image detection is inherently representation-dependent.
- Learned features are manipulation-specific rather than generalisable.
- Detection difficulty varies across manipulation types.
- Robust detection requires the integration of complementary models.

These findings directly validate the thesis's core contributions. First, the empirical characterisation of representation dependency is confirmed by the sharp performance drops observed in cross-dataset evaluations, proving that detection success is a function of feature alignment. Furthermore, the systematic failure mode analysis conducted here reveals that model errors are not random but follow predictable patterns based on the underlying manipulation (e.g., FFT-ResNet failing on spatial CMF). This documentation of structured failures provides the necessary forensic evidence to move beyond "black-box" interpretations. Finally, these insights establish the logical necessity for the representation-aware hierarchical ensemble proposed in the next section. By formally identifying the complementary strengths and non-overlapping failure modes of specialised detectors, this work lays the foundation for an adaptive framework that leverages multiple feature domains to achieve robust, generalisable detection.

6. Hierarchical Ensemble Framework for Robust Fake Image Detection

6.1. Motivation for the Ensemble Framework

The analysis presented in Section 5 suggests a key limitation in fake image detection: model failures do not appear to be arbitrary, but are closely related to the specific feature representations they employ. Models tend to perform well when applied to manipulation types that align with their underlying representation, while their performance degrades when such alignment is lacking[35].

These observations indicate that the challenge is not solely model-specific, but is also influenced by representation-dependent factors. Frequency-based models have been shown to be effective at detecting GAN-induced spectral artefacts[41], yet they are less effective for spatially localised manipulations such as copy-move forgery (CMF). In contrast, models that emphasise spatial similarity are better suited to capturing CMF patterns while being less sensitive to frequency-domain inconsistencies.

Consequently, the observed limitation may not stem from insufficient model capacity, but rather from the constraints of relying on a single representation. This suggests that further improving individual models alone may not be sufficient to achieve robust generalisation across diverse manipulation types.

Instead, a more effective detection system may benefit from integrating multiple complementary representations in a structured manner. However, naïve model aggregation can be insufficient, as it fails to account for the structured nature of model failures and may lead to conflicting or unreliable predictions.

Motivated by these observations, this work introduces a representation-aware hierarchical ensemble framework. Unlike conventional ensembles that treat all models equivalently, the proposed framework employs a conditional decision mechanism to selectively activate models that are more relevant to the input. In this way, the ensemble is formulated as an adaptive system that aligns representation selection with manipulation characteristics, thereby improving detection robustness.

6.2. Proposed Hierarchical Ensemble Framework

A representation-aware hierarchical decision framework is proposed to address the limitations identified in Section 5. Unlike conventional ensemble methods[19], which assume

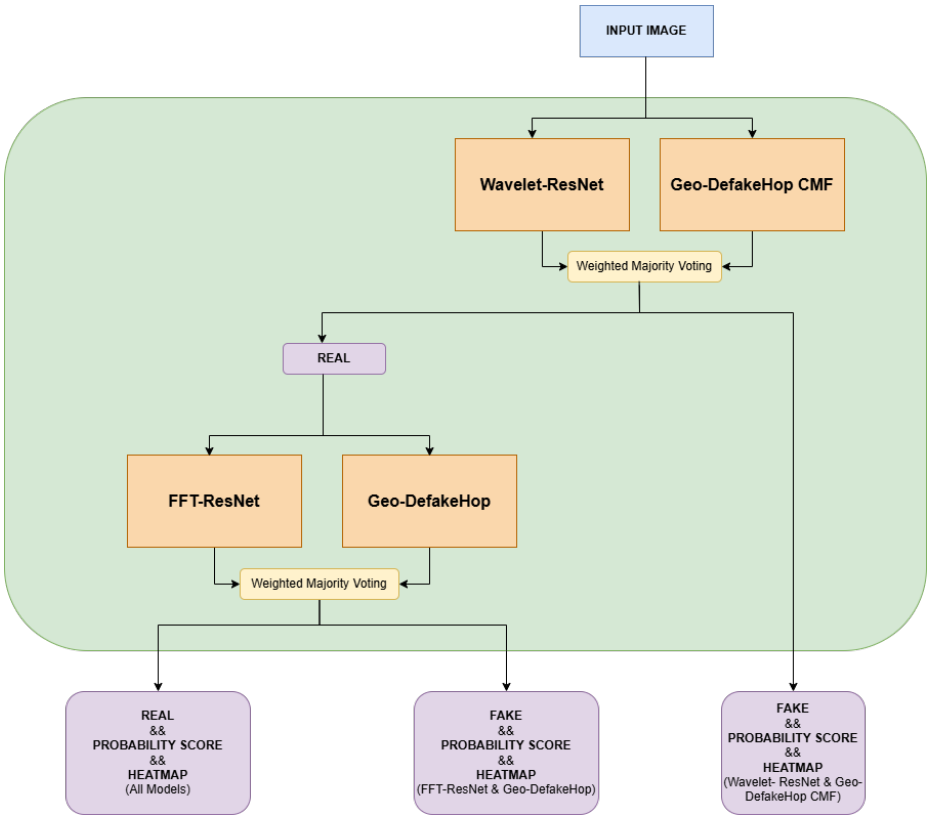


Figure 6.1.: Proposed hierarchical ensemble framework for fake image detection. The system follows a two-stage decision process and generates confidence-weighted, interpretable heatmaps aligned with the final prediction.

all models are equally applicable to all inputs, the proposed system explicitly models the strengths and limitations of each representation. It integrates these within a structured decision process.

Instead of indiscriminately combining models, the framework employs a conditional routing mechanism that selectively activates models based on the input image's characteristics. This design reflects the observation that different manipulation types require different feature representations, and that model failures are structured rather than random. An overview of the proposed architecture is shown in Fig. 6.1.

The framework follows a two-stage hierarchical structure. In the first stage, models specialised in detecting copy-move forgery (CMF) analyse the input image. Specifically, the wavelet-based ResNet and the CMF-trained Geo-DefakeHop model capture spatial inconsistencies such as duplicated regions and local structural similarities[44]. Their predictions are combined using a confidence-weighted voting scheme. Each model contributes to the final prediction in proportion to its confidence.

Prioritising CMF detection in the first stage is motivated by the nature of spatial manipulations. Copy-move forgeries introduce explicit, localised structural cues[21]. These provide strong, directly observable evidence of manipulation. Detecting such patterns early enables the system to confidently identify a subset of fake images without further analysis.

This design choice is further supported by cross-dataset evaluation results. Models trained for CMF detection rely heavily on spatial similarity cues, such as duplicated regions. When applied to GAN-generated images, which typically lack such explicit self-similar structures, these models tend to classify them as real.

This behaviour reflects a representation mismatch and highlights the need to prioritise CMF-specific detection in the first stage, where strong spatial evidence can be reliably identified.

If the image is classified as fake at this stage, the decision is final. Otherwise, the image goes to the second stage, where models specialised in detecting GAN-generated artefacts are applied. In this stage, the FFT-based ResNet and the standard Geo-DefakeHop model identify frequency-domain inconsistencies, particularly in mid- and high-frequency bands. These models are better suited to detecting globally distributed, more subtle artefacts from generative models.

The outputs of the second-stage models are again combined, using the same confidence-weighted voting scheme to produce the final prediction. The weights are determined empirically from validation performance. More reliable models contribute more strongly to the decision process.

This hierarchical design serves as a representation-aware specialist selection mechanism. Rather than treating all models equally, the system ensures each prediction is influenced most by the relevant representation. By filtering out spatially manipulated images in the first stage and analysing spectral inconsistencies only when needed, the framework reduces conflicting model responses and improves consistency.

The proposed framework turns the ensemble from a passive aggregation of models into an adaptive decision system. It explicitly aligns representation selection with manipulation characteristics. This design improves robustness for various manipulation types and directly applies the insights from Section 5.

In addition to the final classification, the proposed framework provides interpretable localisation of manipulation artefacts through hierarchical heatmap generation. Unlike conventional approaches, which produce visual explanations independently of the decision process, the system integrates heatmap generation directly into the hierarchical decision pipeline.

At each stage, model-specific heatmaps are generated using interpretability mechanisms. For ResNet-based models, Grad-CAM[31] is used. For Geo-DefakeHop[4], patch-level probability maps are used. These heatmaps show the spatial distribution of model confidence and highlight regions contributing to the prediction.

The final heatmap is constructed using a confidence-weighted fusion strategy that mirrors the hierarchical decision logic. If a fake prediction is made in the first stage, the final

heatmap is generated by fusing only the CMF-oriented model heatmaps. This ensures the explanation aligns with the detected manipulation type. If the image reaches the second stage, the final heatmap instead comes from GAN-oriented models, capturing frequency-related artefacts.

If both stages contribute to the final decision, heatmaps from each stage are combined using weighted fusion. Each heatmap is scaled by the confidence of its corresponding model. This process ensures that the visual explanation depends directly on the decision signals that determine the final prediction, not on an independent process.

This design establishes a direct link between decision-making and interpretability. Rather than producing generic or post-hoc explanations, the framework generates manipulation-specific heatmaps consistent with the representation and model behaviour. As a result, the system provides accurate predictions and meaningful visual insights into detected artefacts. This close connection between decision-making and explanation further improves reliability and interpretability, making the system particularly suitable for high-stakes applications where both accurate detection and reliable interpretability are critical.

6.3. Experimental Evaluation of the Ensemble System

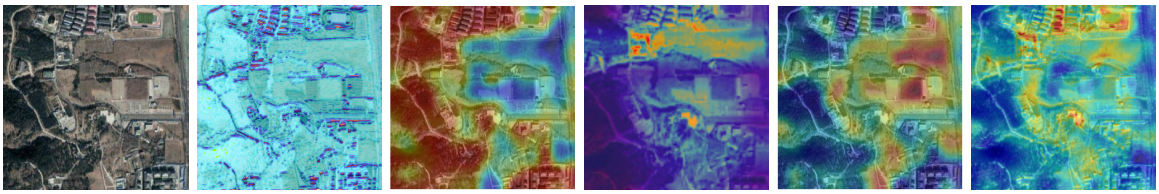
To evaluate the effectiveness of the proposed framework, a mixed test dataset is constructed comprising 400 images, 200 real and 200 fake. The fake images are evenly distributed between GAN-generated images and copy-move forgery (CMF), ensuring that the evaluation reflects realistic scenarios where multiple manipulation types coexist.

Table 6.1.: Performance comparison of individual models and the proposed ensemble system on a mixed dataset containing different types of fake images.

Model	Accuracy	Precision	Recall	F1-score
Geo-DefakeHop	0.645	0.589	**0.960**	**0.730**
Geo-DefakeHop (CMF)	0.558	0.545	0.700	0.613
ResNet (FFT)	0.535	0.529	0.630	0.575
ResNet (Wavelet)	0.513	0.593	0.080	0.141
Proposed Ensemble	**0.715**	**0.787**	0.590	0.674

Table 6.1 presents the performance of individual models and the proposed ensemble framework. The results clearly show that individual models exhibit strong but highly specialised behaviour, whereas the proposed framework achieves more balanced and reliable performance across all manipulation types.

Geo-DefakeHop achieves the highest recall (0.960), indicating a strong ability to correctly identify real images. However, this also suggests a bias towards predicting samples as real. As a result, many manipulated images are misclassified as real, resulting in relatively low precision. This behaviour can be attributed to the model's reliance on global artefact patterns, which may not be consistently present across all types of fake images.



(a) Real Remote Sensing Image (b) FFT-ResNet (c) Geo-DefakeHop (d) Wavelet-ResNet (e) Geo-DefakeHop (CMF) (f) Ensemble Output

Figure 6.2.: Comparison of individual model responses and the final ensemble output, highlighting complementary representation learning across different manipulation characteristics.

In contrast, the wavelet-based ResNet model exhibits the opposite behaviour, with very low recall (0.080), indicating a poor ability to correctly identify real images. This suggests a strong bias towards predicting samples as fake. While its higher precision suggests that its fake predictions are more reliable, the model misclassifies a large number of real images as fake, leading to highly imbalanced performance. This behaviour can be attributed to its strong reliance on spatial duplication cues, which are not present in real images, leading to over-detection of manipulation.

These results highlight a fundamental limitation of individual models: each model is biased towards specific types of artefacts, leading to imbalanced performance when evaluated on heterogeneous datasets[35].

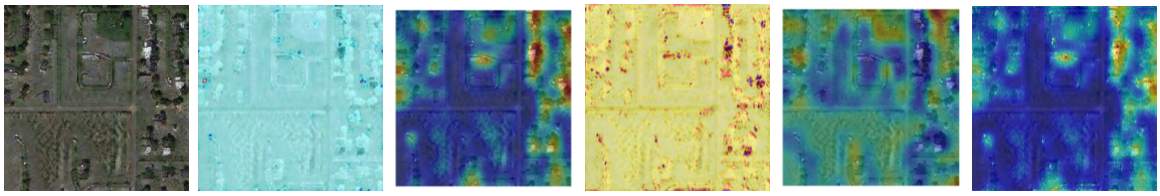
The proposed ensemble framework addresses this limitation by combining complementary model behaviours within a representation-aware decision process. Rather than relying on a single model, the system selectively activates models based on their relevance, balancing detection sensitivity and reliability.

As a result, the proposed ensemble achieves the highest overall accuracy (0.715) and significantly improves precision (0.787) compared to all individual models. Although Geo-DefakeHop attains the highest F1-score (0.730), this is primarily driven by its extremely high recall, which reflects a strong bias towards predicting samples as real.

In contrast, the proposed ensemble provides a more balanced trade-off between precision and recall. While its F1-score is slightly lower, this reflects a more controlled and reliable detection behaviour, avoiding excessive bias towards a single class. In practical scenarios, such balanced performance is more desirable, as it reduces both false positives and false negatives, leading to more stable and trustworthy predictions across heterogeneous manipulation types.

Importantly, the improvement is not only due to model combination, but also due to the alignment between model selection and manipulation characteristics. By ensuring that each decision is primarily influenced by the most appropriate representation, the framework reduces conflicting predictions and improves overall consistency.

In addition to quantitative performance, the behaviour of the proposed framework is further illustrated through qualitative analysis. Figure 6.2, 6.3, 6.4 present example heatmaps generated by individual models and the ensemble for real, CMF and GAN-generated images.

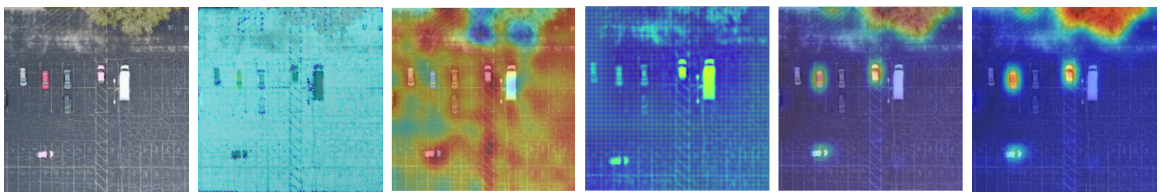


(a) GAN-generated Fake Remote Sensing Image (b) FFT-ResNet (c) Geo-DefakeHop (d) Wavelet-ResNet (e) Geo-DefakeHop (CMF) (f) Ensemble Output

Figure 6.3.: Comparison of individual model responses and the final ensemble output, highlighting complementary representation learning across different manipulation characteristics.

The visual results demonstrate that individual models focus on representation-specific artefacts, often highlighting irrelevant or noisy regions when applied outside their domain of expertise. In contrast, the proposed framework produces more consistent, manipulation-specific heatmaps that reflect the dominant evidence used in the final decision.

This confirms that the proposed system not only improves detection performance but also provides more meaningful and reliable visual explanations. By integrating decision-making and interpretability within a unified framework, the system offers both quantitative and qualitative improvements over individual models.



(a) CMF Fake Remote Sensing Image (b) FFT-ResNet (c) Geo-DefakeHop (d) Wavelet-ResNet (e) Geo-DefakeHop (CMF) (f) Ensemble Output

Figure 6.4.: Comparison of individual model responses and the final ensemble output, highlighting complementary representation learning across different manipulation characteristics.

Despite these improvements, the proposed framework is not without limitations. The performance still depends on the diversity and quality of the underlying models, and the hierarchical decision process may propagate errors if early-stage predictions are incorrect. These limitations will be further discussed in the following section.

Overall, the results demonstrate that fake image detection should be treated as a multi-characteristic problem requiring adaptive analysis. The proposed framework provides a practical and effective solution by leveraging complementary representations within a structured and interpretable decision process.

Ablation Study of the Ensemble Framework

Table 6.2.: Ablation study comparing different ensemble strategies.

Method	Accuracy	Precision	Recall	F1-score
Best Individual Model (Geo-DefakeHop)	0.645	0.589	0.960	0.730
Majority Voting (baseline)	0.565	0.544	0.805	0.649
Proposed Hierarchical Ensemble	**0.715**	**0.787**	0.590	**0.674**

To analyse the contribution of the proposed ensemble design, an ablation study is conducted by comparing it with a baseline ensemble using simple majority voting.

As shown in Table 6.2, the majority voting approach achieves an accuracy of 0.565, which is notably lower than the best individual model (Geo-DefakeHop, 0.645). This result indicates that naive aggregation introduces conflicting decision patterns, as models trained on different representations respond to fundamentally different artefacts. As a result, their predictions are not complementary but often contradictory, resulting in inconsistent decision boundaries.

It is also observed that the majority voting approach increases recall (0.805) but at the cost of precision (0.544), indicating a tendency to over-predict fake samples. In contrast, the proposed framework achieves significantly higher precision, suggesting that the hierarchical selection mechanism effectively suppresses unreliable and misaligned model predictions.

The proposed hierarchical ensemble achieves the highest accuracy of 0.715. This improvement demonstrates that performance gains are not due only to model combination, but to the representation-aware decision mechanism, which selectively utilises models based on their strengths.

Furthermore, the proposed system achieves higher precision than all baselines, indicating more reliable fake image detection while maintaining a balanced trade-off with recall. These findings highlight that performance improvement is not an inherent property of ensemble learning, but depends on how model outputs are integrated. Simply combining multiple models without considering their representational compatibility can lead to performance degradation rather than improvement.

Overall, the ablation study confirms that the effectiveness of the proposed framework stems from its representation-aware design rather than the mere combination of multiple models. By explicitly aligning model selection with manipulation-specific characteristics, the framework mitigates representation conflicts and achieves more reliable and generalisable detection performance. This behaviour further supports the central hypothesis of this thesis: fake image detection is inherently representation-dependent, and effective solutions must explicitly account for this dependency.

Part V.

Conclusion and Future Work

7. Conclusion and Future Work

7.1. Conclusion

This thesis establishes that fake remote sensing image detection is fundamentally a representation-dependent problem. Through systematic experimental analysis across GAN-generated and copy-move forgery (CMF) datasets, it is shown that different manipulation types introduce distinct and non-overlapping characteristics that may not be effectively captured by a single model or feature representation.

The results reveal that detection models struggle to learn universal forgery patterns and instead develop representation-specific sensitivities. The analysis shows that frequency-based approaches are highly effective for detecting GAN-generated images with global spectral inconsistencies, whereas spatially localised representations are more suitable for identifying CMF, where manipulation manifests as structural duplication. However, these representations fail to generalise across manipulation types, leading to consistent, structured performance degradation across datasets and manipulation types.

Importantly, this limitation cannot be attributed solely to insufficient model capacity. Instead, it arises from a fundamental mismatch between the chosen feature representation and the underlying manipulation characteristics, rather than a limitation of model capacity. This finding challenges the assumption that a single unified model can robustly detect all types of fake images.

Building upon this insight, a representation-aware hierarchical ensemble framework is proposed. Unlike conventional ensemble methods, which treat all models equally, the proposed approach explicitly aligns model selection with manipulation-specific characteristics through a conditional decision mechanism. Experimental results show that naive ensemble strategies, such as majority voting, fail to improve performance and may even degrade it due to conflicting predictions. In contrast, the proposed framework achieves more reliable and balanced performance by leveraging complementary model behaviours in a structured manner. These results demonstrate that effective ensemble design in fake image detection should be representation-aware rather than model-agnostic.

These findings suggest that fake image detection should be reformulated as a multi-characteristic problem rather than a single-domain classification task. Effective solutions must explicitly account for the diversity of manipulation artefacts and the representation-dependent nature of detection models.

Overall, this work demonstrates that the limitations of existing fake image detection methods stem from their reliance on single representations and establishes that robust detection requires adaptive, representation-aware strategies that explicitly integrate com-

plementary feature domains. This perspective provides a foundation for future research on more generalisable and interpretable detection systems in remote sensing and beyond.

7.2. Future Work

While this thesis demonstrates that fake image detection is inherently representation-dependent and proposes an effective ensemble framework, several directions remain for further improvement and extension.

First, the current study focuses on a limited set of manipulation types, primarily GAN-generated images and copy-move forgery (CMF). Future work could extend this framework to include additional manipulation types, such as splicing, inpainting, or diffusion-based image generation. Incorporating a broader range of forgery mechanisms would enable a more comprehensive evaluation of representation-dependent behaviour.

Second, although the proposed ensemble system improves robustness, its decision fusion strategy relies on predefined rules and thresholds. Future research could explore learnable or adaptive fusion mechanisms, such as attention-based weighting or meta-learning approaches, to dynamically optimise model contributions based on input characteristics.

Third, the current models operate on fixed input resolutions and predefined transformations (e.g., FFT and wavelets). Future work could investigate multi-scale and multi-resolution representations, as well as hybrid feature extraction techniques that jointly learn spatial and frequency-domain information in an end-to-end manner.

Another important direction is improving generalisation across unseen datasets and manipulation types. Domain adaptation and self-supervised learning techniques could be explored to reduce dependency on labelled data and enhance robustness to distribution shifts.

In addition, while Geo-DefakeHop provides interpretability through channel-wise analysis, integrating interpretability directly into deep learning models remains an open challenge. Future work could focus on developing inherently interpretable architectures that combine the transparency of statistical methods with the flexibility of deep learning.

Finally, the current evaluation is conducted under controlled experimental settings. Extending the framework to real-world scenarios, including high-resolution satellite imagery and complex environmental conditions, would further validate its practical applicability.

Overall, these directions aim to develop a more general, adaptive, and scalable framework for detecting fake remote sensing images.

Appendix

A. Implementation Details

A.1. FFT-based ResNet Model for GAN-generated Fake Images

A.1.1. FFT-based Input Representation

To explicitly capture frequency-domain artefacts introduced by generative models, the input images are transformed into the frequency domain using the 2D Fast Fourier Transform (FFT).

Given an input image $I(x, y)$ of spatial size $H \times W$, its 2D discrete Fourier transform is defined as:

$$F(u, v) = \sum_{x=0}^{H-1} \sum_{y=0}^{W-1} I(x, y) \cdot e^{-j2\pi\left(\frac{ux}{H} + \frac{vy}{W}\right)} \tag{A.1}$$

where (u, v) denote frequency coordinates and j is the imaginary unit.

Since the Fourier transform produces complex-valued outputs, we compute the magnitude spectrum:

$$M(u, v) = |F(u, v)| = \sqrt{\Re(F(u, v))^2 + \Im(F(u, v))^2} \tag{A.2}$$

To stabilise the dynamic range and enhance the visibility of high-frequency components, logarithmic scaling is applied:

$$\hat{M}(u, v) = \log(1 + M(u, v)) \tag{A.3}$$

Finally, the spectrum is shifted such that the zero-frequency component is centred:

$$M_{shifted} = \text{FFTShift}(\hat{M}) \tag{A.4}$$

This frequency representation is used as the input to the CNN model instead of the raw RGB image.

A.1.2. Model Architecture

The FFT-based representation is fed into a convolutional neural network utilising the ResNet architecture, adapted here to operate on frequency-domain inputs instead of spatial RGB images. The specific configuration of this FFT-based ResNet model is detailed below.

Table A.1.: Architecture of FFT-based ResNet Model

Component	Configuration
Input	Log-scaled high-pass filtered FFT magnitude spectrum
Input Size	$H \times W \times C$
Backbone	ResNet-34
First Conv Layer	7×7, stride 2
Pooling	MaxPooling (3×3)
Residual Blocks	Standard ResNet blocks
Activation	ReLU
Normalization	Batch Normalization
Global Pooling	Adaptive Average Pooling
Fully Connected	2 output neurons
Output	Softmax (binary classification)

A.1.3. Training Configuration

The model is trained using the following hyperparameters:

Table A.2.: Training Hyperparameters

Parameter	Value
Optimizer	SGD
Learning Rate	0.01
Batch Size	16
Epochs	20
Loss Function	Binary Cross-Entropy
Input Size	224×224

The Adam optimizer and the learning rate is selected based on paper. The model is trained for a fixed number of epochs.

A.1.4. Inverse FFT for Spatial Visualization

Since the proposed model operates on frequency-domain representations, the resulting Grad-CAM activation maps are also obtained in the frequency domain. To interpret these activations in the spatial domain, an inverse Fourier transform is applied.

Given the frequency-domain activation map $L_{\text{freq}}(u, v)$, the corresponding spatial-domain representation is obtained via the inverse 2D Fourier transform:

$$L_{\text{spatial}}(x, y) = \frac{1}{HW} \sum_{u=0}^{H-1} \sum_{v=0}^{W-1} L_{\text{freq}}(u, v) \cdot e^{j2\pi\left(\frac{ux}{H} + \frac{vy}{W}\right)} \tag{A.5}$$

where (u, v) denote frequency coordinates and (x, y) denote spatial coordinates.

Since the activation map is real-valued in practice, only the real component of the inverse transform is retained:

$$\hat{L} * \text{spatial}(x, y) = \Re\left(L * \text{spatial}(x, y)\right) \tag{A.6}$$

Optionally, the resulting spatial map is normalized to the range $[0, 1]$ for visualization purposes:

$$L_{\text{norm}} = \frac{\hat{L}_{\text{spatial}} - \min}{\max - \min} \tag{A.7}$$

This transformation enables the interpretation of frequency-domain activations in the original image space, allowing direct comparison with spatial structures and visual artefacts.

A.2. Wavelet-based ResNet Model for Copy-Move Forgery Detection

A.2.1. Wavelet-based Input Representation

To capture spatial inconsistencies and localised artefacts introduced by copy-move forgeries, the input images are transformed using the Discrete Wavelet Transform (DWT) with Haar wavelets, which provides a multi-scale representation of the image.

Given an input image $I(x, y)$, a single-level 2D DWT decomposes it into four sub-bands:

$$LL, LH, HL, HH = \text{DWT}(I) \tag{A.8}$$

where:

- LL represents the low-frequency approximation component,
- LH captures horizontal edge information,
- HL captures vertical edge information,
- HH captures diagonal high-frequency details.

These sub-bands encode both coarse structures and fine-grained local variations, which are essential for detecting duplicated regions in copy-move forgery.

The sub-bands are concatenated along the channel dimension to form the input tensor:

$$I_{\text{wavelet}} = \text{Concat}(LL, LH, HL, HH) \tag{A.9}$$

This multi-channel representation is then used as input to the CNN model.

A.2.2. Model Architecture

The wavelet-based representation is processed using a ResNet architecture adapted for multi-channel input. The model structure remains similar to the FFT-based variant, with modifications in the input layer to accommodate wavelet sub-bands.

Table A.3.: Architecture of Wavelet-based ResNet Model

Component	Configuration
Input	Wavelet sub-bands (LL, LH, HL, HH)
Input Size	$H \times W \times 4C$
Backbone	ResNet-34
First Conv Layer	7×7, stride 2
Pooling	MaxPooling (3×3)
Residual Blocks	Standard ResNet blocks
Activation	ReLU
Normalization	Batch Normalization
Global Pooling	Adaptive Average Pooling
Fully Connected	2 output neurons
Output	Softmax (binary classification)

A.2.3. Training Configuration

The training configuration follows the same setup as the FFT-based model for consistency.

Table A.4.: Training Hyperparameters

Parameter	Value
Optimizer	SGD
Learning Rate	0.01
Batch Size	16
Epochs	20
Loss Function	Binary Cross-Entropy
Input Size	224×224

Unlike FFT which provides global frequency information, DWT preserves spatial locality, making it more suitable for detecting localised manipulations such as copy-move forgery.

A.3. Geo-DefakeHop Implementation Details

The Geo-DefakeHop model is implemented with the following configuration.

Input Preprocessing:

- Input images are resized to 256×256
- Patch size: 16×16
- Number of patches per image: 256 (non-overlapping grid)
- Color space: RGB

Patch extraction is performed using a fixed grid structure, where each image is divided into 16×16 patches without overlap. Only images producing exactly 256 patches are retained.

Multi-hop Saab Configuration:

- Number of hops: 3 (HopA, HopB, HopC)
- Kernel sizes: [2, 3, 4]
- Split threshold: 0.01
- Energy threshold (keep_thr): 0.001

At each hop, the number of output channels is determined based on energy compaction, where components with eigenvalue energy above a threshold are retained.

Channel Selection:

- Channel-wise classifiers are trained for each hop
- Channels are selected based on F1-score

Feature Aggregation:

- Patch-level features are extracted independently
- Features are aggregated to image-level by concatenation
- Patch-level channel-wise probabilities are concatenated to form image-level feature vectors

Final Classifier:

- Model: XGBoost
- Number of estimators: 100
- Learning rate: 0.2
- Max depth: 6

- Objective: binary logistic
- Evaluation metric: AUC

The class imbalance is handled using a dynamic `scale_pos_weight` parameter based on the ratio of fake and real samples.

Decision Rule:

- Final prediction is obtained using a probability threshold of 0.5

The same configuration is applied across datasets for consistency.

B. Ensemble Model Implementation Details

B.1. Confidence-Weighted Decision Strategy

The ensemble framework employs a confidence-weighted voting mechanism to combine predictions from multiple models.

Each model produces a predicted label $y_i \in \{\text{Real}, \text{Fake}\}$ and a confidence score $c_i \in [0, 1]$. To unify predictions, labels are mapped to signed scores:

$$s_i = \begin{cases} +c_i, & \text{if } y_i = \text{Real} \\ -c_i, & \text{if } y_i = \text{Fake} \end{cases} \tag{B.1}$$

The final decision score is computed as:

$$S = \sum_{i=1}^{N} s_i \tag{B.2}$$

The final prediction is defined as:

$$\hat{y} = \begin{cases} \text{Real}, & \text{if } S \geq 0 \\ \text{Fake}, & \text{if } S < 0 \end{cases} \tag{B.3}$$

The corresponding confidence is given by:

$$C = \frac{|S|}{N} \tag{B.4}$$

This formulation ensures that both prediction and confidence reflect model agreement and reliability.

B.2. Hierarchical Decision Flow

The ensemble follows a two-stage hierarchical decision process.

Stage 1: CMF-specialised models

The first stage combines predictions from CMF-oriented models:

$$S_1 = s_{\text{geo\_cmf}} + s_{\text{resnet\_wavelet}} \tag{B.5}$$

If:

$$S_1 < 0 \Rightarrow \text{Final = Fake} \tag{B.6}$$

Otherwise, the sample proceeds to Stage 2.

Stage 2: GAN-specialised models

$$S_2 = s_{\text{geo\_gan}} + s_{\text{resnet}} \tag{B.7}$$

The decision is defined as:

$$\hat{y} = \begin{cases} \text{Fake}, & \text{if } S_2 < 0 \\ \text{Real}, & \text{otherwise} \end{cases} \tag{B.8}$$

Final Aggregation

If all models indicate real samples, a final aggregation is performed:

$$S_{\text{final}} = \sum_{i=1}^{4} s_i \tag{B.9}$$

$$C_{\text{final}} = \frac{|S_{\text{final}}|}{4} \tag{B.10}$$

B.3. Decision Flow Pseudocode

```
function hierarchical_ensemble(models):

    # Stage 1: CMF models
    score_stage1 = vote(geo_cmf) + vote(resnet_wavelet)

    if score_stage1 < 0:
        return Fake, abs(score_stage1) / 2

    # Stage 2: GAN models
    score_stage2 = vote(geo_gan) + vote(resnet)

    if score_stage2 < 0:
        return Fake, abs(score_stage2) / 2

    # Final aggregation
    total_score = sum(all model votes)
```

```
if total_score >= 0:
    return Real, abs(total_score) / 4
else:
    return Fake, abs(total_score) / 4
```

B.4. Confidence-Weighted Heatmap Fusion

To generate interpretable outputs, model-specific heatmaps are fused using confidence-weighted averaging.

Given two heatmaps H_1, H_2 and confidences c_1, c_2, the fused heatmap is defined as:

$$H_{\text{fused}} = \frac{c_1 \cdot H_1 + c_2 \cdot H_2}{c_1 + c_2 + \epsilon} \tag{B.11}$$

Each heatmap is first normalised:

$$H = \frac{H - \min(H)}{\max(H) - \min(H) + \epsilon} \tag{B.12}$$

The final heatmap is also normalised:

$$H_{\text{final}} = \frac{H_{\text{fused}} - \min}{\max - \min + \epsilon} \tag{B.13}$$

This formulation ensures that higher-confidence models contribute more strongly to the final visual explanation.

B.5. Implementation Details

The ensemble system is implemented with the following configuration:

- Heatmap resolution: 224×224
- Data type: float32
- Patch size (Geo-DefakeHop): 16×16
- Number of models: 4
- Fusion stages:
  - CMF heatmap fusion
  - GAN heatmap fusion
  - Final-stage fusion

Bibliography

[1] Younis Abdalla, M. Tariq Iqbal, and Mohamed Shehata. Copy-move forgery detection and localization using a generative adversarial network and convolutional neural-network. *Information*, 10(9), 2019.

[2] Ritu Agarwal, Deepak Khudaniya, Abhinav Gupta, and Khyati Grover. Image forgery detection and deep learning techniques: A review. In *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 1096–1100, 2020.

[3] Anindya Bhattacharjee, Kaidul Islam, Kafi Anan, Ashir Intesher, Abrar Assaeem Fuad, Utsab Saha, and Hafiz Imtiaz. Cae-net: Generalized deepfake image detection using convolution and attention mechanisms with spatial and frequency domain features. *Journal of Visual Communication and Image Representation*, 115:104679, 2026.

[4] Hong-Shuo Chen, Kaitai Zhang, Shuowen Hu, Suya You, and C.-C. Jay Kuo. Geodefakehop: High-performance geographic fake image detection. *APSIPA Transactions on Signal and Information Processing*, 13, 01 2024.

[5] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[6] Wei Chen, Jiage Chen, Yuewu Wan, Xining Liu, Mengya Cai, Jingguo Xu, Hongbo Cui, and Mengdie Duan. Land cover classification based on multimodal remote sensing fusion. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, X-1-2024:35–40, 05 2024.

[7] Yueru Chen and C-C Jay Kuo. Pixelhop: A successive subspace learning (ssl) method for object recognition. *Journal of Visual Communication and Image Representation*, 70:102749, 2020.

[8] Vincent Christlein, Christian Riess, Johannes Jordan, Corinna Riess, and Elli Angelopoulou. An evaluation of popular copy-move forgery detection approaches. *IEEE Transactions on information forensics and security*, 7(6):1841–1854, 2012.

[9] Umur Aybars Ciftci and Ilke Demir. Deepfake satellite imagery detection with multi-attention and super resolution. *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, pages 4871–4874, 2023.

[10] Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. Efficient dense-field copy–move forgery detection. *IEEE Transactions on Information Forensics and Security*, 10(11):2284–2297, 2015.

[11] Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection. In *Proceedings of the 5th ACM workshop on information hiding and multimedia security*, pages 159–164, 2017.

[12] Tarik Dzanic, Karan Shah, and Freddie Witherden. Fourier spectrum discrepancies in deep network generated images. *Advances in neural information processing systems*, 33:3022–3032, 2020.

[13] M. V. Gashnikov and A. V. Kuznetsov. Detection of fake remote-sensing data. *Opt. Mem. Neural Netw.*, 31(1):16–21, March 2022.

[14] János Horváth, Daniel Mas Montserrat, Hanxiang Hao, and Edward J. Delp. Manipulation detection in satellite images using deep belief networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2832–2840, 2020.

[15] Hailing Huang, Weiqiang Guo, and Yu Zhang. Detection of copy-move forgery in digital images using sift algorithm. In *2008 IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application*, volume 2, pages 272–276, 2008.

[16] Guonian Jin and Xiaoxia Wan. An improved method for sift-based copy–move forgery detection using non-maximum value suppression and optimized j-linkage. *Signal Processing: Image Communication*, 57:113–125, 2017.

[17] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan, 12 2019.

[18] Harpreet Kaur, Jyoti Saxena, and Sukhjinder Singh. Key-point based copy-move forgery detection and their hybrid methods: A review. *Journal of Advance Research in Electrical & Electronics Engineering (ISSN: 2208-2395)*, 2:06–12, 06 2015.

[19] Ludmila Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms: Second Edition*, volume 47. 01 2014.

[20] Guohui Li, Qiong Wu, Dan Tu, and ShaoJie Sun. A sorted neighborhood approach for detecting duplicated regions in image forgeries based on dwt and svd. pages 1750–1753, 07 2007.

[21] Yuanman Li, Yingjie He, Changsheng Chen, Li Dong, Bin Li, Jiantao Zhou, and Xia Li. Image copy-move forgery detection via deep patchmatch and pairwise ranking learning. *IEEE Transactions on Image Processing*, 34:425–440, 2025.

[22] Arpan Mahara and Naphtali Rishe. Methods and trends in detecting ai-generated images: A comprehensive review. *Computer Science Review*, 60:100908, 2026.

[23] S.G. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989.

[24] S.G. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989.

[25] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do gans leave artificial fingerprints? In *2019 IEEE conference on multimedia information processing and retrieval (MIPR)*, pages 506–511. IEEE, 2019.

[26] Ghulam Muhammad, Muhammad Hussain, and George Bebis. Passive copy move image forgery detection using undecimated dyadic wavelet transform. *Digital Investigation*, 9(1):49–57, 2012.

[27] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 1, 10 2016.

[28] Ji Qi, Xinchang Zhang, Dingqi Ye, Ruan Yongjian, Xin Guo, Shaowen Wang, and Haifeng Li. Sfnet: Fusion of spatial and frequency-domain features for remote sensing image forgery detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, PP:1–16, 01 2025.

[29] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International conference on machine learning*, pages 5301–5310. PMLR, 2019.

[30] Tamer Say, Mustafa Alkan, and Aynur Kocak. Advancing gan deepfake detection: Mixed datasets and comprehensive artifact analysis. *Applied Sciences*, 15(2), 2025.

[31] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: visual explanations from deep networks via gradient-based localization. *International journal of computer vision*, 128(2):336–359, 2020.

[32] Eero P. Simoncelli and Bruno A. Olshausen. Natural image statistics and neural representation. *Annual review of neuroscience*, 24:1193–216, 2001.

[33] Jialu Sui, Ding Ma, C.-C. Jay Kuo, and Man-On Pun. Fldcf: A collaborative framework for forgery localization and detection in satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024.

[34] Hang Tu, Peng Liang, Xiaoguang Lu, and Huimin Zhao. Mgcfdn: Image copy-move forgery detection method based on multi-granularity feature consistency. *Neurocomputing*, 664:132029, 11 2025.

[35] Luisa Verdoliva. Media forensics and deepfakes: An overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):910–932, 2020.

[36] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8695–8704, 2020.

[37] Shaowei Weng, Tangguo Zhu, Tiancong Zhang, and Chunyu Zhang. Ucm-net: A u-net-like tampered-region-related framework for copy-move forgery detection. *IEEE Transactions on Multimedia*, 26:750–763, 2024.

[38] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3974–3983, 2018.

[39] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

[40] Xiuxiao Yuan, Shiyu Chen, Wei Yuan, and Yang Cai. Poor textural image tie point matching via graph theory. *ISPRS Journal of Photogrammetry and Remote Sensing*, 129:21–31, 2017.

[41] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. In *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6, 2019.

[42] Ze Zhang, Enyuan Zhao, Yi Jiang, Nie Jie, and Xinyue Liang. Challenging dataset and multi-modal gated mixture of experts model for remote sensing copy-move forgery understanding. pages 1–6, 06 2025.

[43] Bo Zhao, Shaozeng Zhang, Chunxue Xu, Yifan Sun, and Chengbin Deng. Deep fake geography? when geospatial data encounter artificial intelligence. *Cartography and Geographic Information Science*, 48(4):338–352, 2021.

[44] Jiangbin Zheng, Yanan Liu, Jinchang Ren, Tingger Zhu, Yijun Yan, and Heng Yang. Fusion of block and keypoints based approaches for effective copy-move image forgery detection. *Multidimensional Systems and Signal Processing*, 27, 10 2016.

[45] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. pages 2242–2251, 10 2017.