

Forensic Analysis and Detection of Manipulated Images Using Deep Learning Techniques

by Sakshi Final

Submission date: 27-May-2026 02:10PM (UTC+0530)

Submission ID: 2970462815

File name: thesis_6.pdf (503.44K)

Word count: 11622

Character count: 68721

CONTENTS

Abstract	v
1 Introduction	1
1.1 Image Manipulation Operations and Their Forensic Signatures	1
1.2 JPEG Compression and the Forensic Detection Challenge	2
1.3 Limitation of Equal Feature Weighting in Dual-Stream Systems . . .	2
1.4 Attention Mechanisms as a Principled Solution	3
1.5 Problem Statement	3
1.6 Objectives of the Study	4
1.7 Organisation of the Thesis	5
2 Technical Background	6
2.1 Image Manipulation Operations and Forensic Fingerprints	6
2.2 JPEG Compression and Its Forensic Impact	7
2.3 Deep Learning Building Blocks	7
2.3.1 Convolutional Neural Networks	8
2.3.2 Residual Networks	8
2.3.3 DCT and Frequency-Domain Analysis	8
2.3.4 Attention Mechanisms: CBAM	9
2.4 Dataset, Manipulation Categories, and Resolution Analysis	9
2.4.1 Dataset Construction	9
2.4.2 Manipulation Categories	10
2.4.3 Image Resolution and Forensic Implications	10
3 Related Work	11
3.1 Traditional Forensic Detection Methods	11
3.2 Deep Learning Approaches to Manipulation Detection	12
3.3 Manipulation Chain Detection	13
3.4 Attention Mechanisms in Image Forensics	13
3.5 Research Gap and Motivation	14

4	Proposed Methodology	16
4.1	Overview of the AFNet Architecture	16
4.2	Noise Residual Attention Stream	16
4.3	Compression Artifact Stream	17
4.4	CBAM Module	18
4.5	Multi-Point CBAM Integration Strategy	19
4.5.1	CBAM within Residual Blocks	19
4.5.2	CBAM after Initial Convolutional Layers	20
4.5.3	CBAM within the Cross-Stream Fusion Layer	20
4.6	Complete Architecture and Classification Head	22
4.7	Training Strategy	22
4.8	Dataset Construction and Experimental Setup	22
5	Experimental Results and Evaluation	24
5.1	Overall Performance on RAISE ALL	24
5.2	Per-Class Performance Analysis	25
5.3	Ablation Studies	26
5.4	Cross-Dataset Generalisation	28
5.5	Computational Overhead Analysis	29
6	Discussion and Analysis	30
6.1	Architectural Development and the Role of Each Design Decision	30
6.2	Channel and Spatial Attention in the Forensic Context	31
6.3	The Compound Smoothing Chain Detection Problem	31
6.4	Generalisation and Domain Invariance	32
6.5	Limitations	32
7	Conclusion and Future Work	33
7.1	Conclusion	33
7.2	Future Work	34

LIST OF FIGURES

4.1	AFNet dual-stream architecture with NRAS and CAS streams, DCT transformation, CBAM-enhanced fusion, and classification head.	17
4.2	NRAS stream with convolutional layers, regular residual blocks (cyan), and pooling residual blocks (orange).	17
4.3	CAS stream with DCT transformation, parallel dilated convolutions, and residual blocks.	18
4.4	CBAM Channel Attention Module: uses global average pooling and global max pooling, processed through a shared MLP with bottleneck ratio $r = 16$, and sigmoid activation to produce the channel attention map.	19
4.5	CBAM Spatial Attention Module: uses channel-wise average pooling and max pooling, concatenation, 7×7 convolution, and sigmoid activation to produce the spatial attention map.	20
4.6	Regular residual block with CBAM (magenta) after the second convolution and before the skip addition.	21
4.7	Pooling residual block with CBAM and max pooling (red) on the skip path for spatial downsampling.	21

LIST OF TABLES

3.1 Comparison of Major Image Forensic Detection Approaches.	15
5.1 Overall Accuracy Comparison on RAISE ALL (%).	24
5.2 Per-Class Performance Comparison on RAISE ALL (%).	25
5.3 Ablation Study: Effect of Attention Placement and Formulation on Validation and Test Accuracy (%).	27
5.4 Cross-Dataset Evaluation: Accuracy on DRESDEN and UCID (%).	28

ABSTRACT

The detection of image manipulation chains in JPEG-compressed environments remains among the most consequential and technically demanding problems in the field of digital image forensics. Real-world forgeries rarely involve isolated, single-step processing operations; instead, they are constructed through ordered sequences of manipulations, spatial filtering, blurring, resampling, and noise injection which are subsequently subjected to lossy JPEG compression prior to distribution. This compound character of practical forgeries fundamentally undermines single-operation forensic detectors, which perform well in controlled settings but fail to generalise when successive operations systematically obscure earlier forensic evidence.

This thesis presents AFNet (AttentionForensicNet), a novel attention-driven dual-stream convolutional network that addresses the combined challenges of manipulation chain detection and lossy compression through strategic integration of the Convolutional Block Attention Module (CBAM) within a dual-stream forensic framework. AFNet processes 256×256 grayscale image patches through two dedicated streams: the Noise Residual Attention Stream (NRAS), which extracts spatial-domain forensic fingerprints through residual learning, and the Compression Artifact Stream (CAS), which operates in the DCT frequency domain to model JPEG-specific artefacts and their interaction with prior manipulation traces. CBAM is integrated at three complementary locations, within every residual block of both streams, after the initial convolutional layers of each stream, and within the cross-stream fusion layer, enabling the network to perform dynamic, content-aware feature selection at each stage of the processing hierarchy.

Comprehensive experimental evaluation on the RAISE ALL benchmark demonstrates that AFNet achieves 97.17% overall test accuracy, surpassing all evaluated baseline systems including constrained CNNs, multi-scale residual networks, and prior dual-stream architectures. Per-class analysis reveals especially pronounced improvements on the hardest compound smoothing chains: 97.79% on MF→GB chains and 92.15% on GB→MF chains, compared to 84.10% achieved by the strongest single-stream competitor. Cross-dataset evaluation on the DRESDEN and UCID benchmarks confirms that the attention-learned features generalise across diverse camera systems and acquisition conditions. Ablation studies validate that every CBAM integration point contributes uniquely, and that joint channel-spatial attention is superior to channel-only or spatial-only alternatives. The complete attention augmentation introduces only 0.08M additional parameters and approximately 2.2% additional inference latency, establishing AFNet as both highly effective and computationally practical.

CHAPTER 1

INTRODUCTION

The deep penetration of digital imaging technologies into the areas where their impact is most critical—legal matters, journalism, medicine, intelligence, and social media—raises serious concerns about the need to verify the authenticity of visual images reliably using computation [2]. With advanced image editing programs and sophisticated AI-enabled content-generation systems, the technological hurdles for forgery have been drastically reduced, putting enormous strain on content verification pipelines [3]. The stakes of forensic failure can be devastating: fake photos can influence court cases, manipulated medical imagery leads to incorrect diagnosis, and fabricated news photos can disseminate fake news at large scales.

The study of digital image forensics has become the scientific field responsible for the detection and analysis of alterations to images [4]. Ideally, a forensic technique is supposed to do more than simply ascertain whether there have been any alterations made to an image; it should be able to trace all the processing steps used on the image, including the nature of each step, its parameters, and the order of steps. This type of investigation, referred to as manipulation chain detection, is much more difficult than mere operation detection and forms the crux of this our work.

1.1 Image Manipulation Operations and Their Forensic Signatures

Every typical operation on the spatial domain used in forensic analyses causes an identifiable and analyzable change to the signal of the image. Median filtering is often applied for impulse noise removal, and the impact it makes on the correlation structure of pixels in a photograph is distinctively different from the normal pattern, causing deviations within the noise residual spectrum [5]. Applying Gaussian blurring will cause certain low-pass effects in the residual domain, since Gaussian blurring averages out the high frequency signals of the pixel matrix [6]. Resampling, scaling, and rotation of an image will cause periodic distortions to the pixel matrix due to interpolation techniques; hence, creating distinct peaks in the spectrum at identifiable frequencies [7]. The process of contrast enhancement changes the distribution of pixels in an image,

breaking the camera response model [8].

Recent advances in the construction of one-shot fingerprinting algorithms allow identifying these separate forensic fingerprints with good accuracy in a laboratory setting [2, 3]. What makes the problem difficult is that operations are executed sequentially. When combining two operations, we do not get the sum of their fingerprints – we get something entirely new and much more complex. Notably, using median filtering and Gaussian blur in sequence will leave a significantly different forensic footprint compared to reversing their order, and both of these footprints will be completely distinct from the footprints left by each operation separately [2, 4]. Order-insensitive systems will fail to recognize such cases consistently, making the design of order-sensitive systems necessary.

1.2 JPEG Compression and the Forensic Detection Challenge

In virtually all real-world forensic scenarios, manipulated images are ultimately subjected to JPEG compression before distribution, through social media upload, email transmission, or consumer device storage. JPEG compression introduces its own family of artefacts: blocking discontinuities at 8×8 block boundaries in the spatial domain, and quantization noise in the DCT coefficient domain. These compression-induced perturbations interact with and partially mask the forensic traces of prior spatial-domain manipulations, because quantization noise can modify the subtle statistical anomalies left by earlier operations [4, 9]. This interaction is the primary reason why effective forensic chain detection under real-world conditions demands dedicated compression-aware architectures capable of explicitly modeling DCT-domain behaviour while amplifying the most informative forensic evidence [9, 10].

1.3 Limitation of Equal Feature Weighting in Dual-Stream Systems

The dual-stream architecture, in which separate network branches process spatial-domain and frequency-domain representations in parallel before combining their features for classification, has emerged as the dominant design paradigm for compression-aware forensic chain detection [4, 10]. The pioneering research by Kadha et al. [9] illustrated how a combination of a noise residual extraction pipeline in the spatial domain and a compression feature extraction pipeline in the DCT domain results in significantly improved accuracy compared to both approaches used separately, achieving an accuracy of over 91.98%. Despite the remarkable results achieved with the dual stream method, one significant the equality of the importance assigned to each ex-

tracted channel and every position in space when extracting and combining features. In forensic studies, the importance assigned to each feature equally is certainly not optimal. The forensics information that would help distinguish the two manipulation chains does not exist equally among the channels and spatial positions; some frequency ranges, some spatial locations, and some neural network layers have significantly more discriminative ability [11] [12].

1.4 Attention Mechanisms as a Principled Solution

Attention mechanisms offer a compelling computational model for solving the problem of feature weighting. By training to give varying degrees of importance to each of the feature channels and spatial locations, the attention mechanism helps allocate the representational capabilities of the network towards the salient aspects of the feature representation. The Convolutional Block Attention Module (CBAM) [1], which employs sequential application of channel and spatial attention, has shown consistent performance improvements in different computer vision applications. The successful use of CBAM in forensic analysis applications such as image splicing detection [11] and copy-move forgery detection [12] corroborates the transferability of this feature selection mechanism to the domain of forensic analysis.

The key idea underlying our work is that the insertion of CBAM in a number of appropriately selected locations within a dual-stream architecture can help the network deal with compression noise and extract the salient discriminating characteristics in a way that results in significant improvements in detection accuracy of compound manipulation chains, which pose a severe challenge for non-adaptive algorithms.

1.5 Problem Statement

As digital images become easier to manipulate, verifying their authenticity has become one of the most pressing challenges in forensic research. Current detection methods struggle to keep pace with modern editing tools, and the following core problems remain largely unsolved:

1. **Vanishing Forensic Traces:** When an image is processed using common operations such as median filtering, it leaves behind faint statistical fingerprints. However, these traces are fragile any subsequent editing step can quietly erase or disguise them, leaving investigators with nothing to analyze.
2. **Masking Effect of JPEG Compression:** Almost every picture uploaded online is subject to compression using JPEG. The compression algorithm creates its

own noise, adding to the already existing artifacts and thus masking them in such a way that false alarms become commonplace.

3. **Masking Due to Multi-step Image Manipulation:** Editing an image usually includes cropping, resizing, applying filters, and then saving the result. During each of the following steps, any traces left behind during a previous step are covered up by another one. Hence, it becomes hard to recreate the whole history of manipulations done to an image file.
4. **Limitations due to Low Resolution Images:** Almost all existing forensic methods were created using images with high resolution. As soon as a high-resolution image is scaled down, any features it used to contain become less noticeable and even trained classifiers fail to distinguish between real textures and fabricated ones.

1.6 Objectives of the Study

¹² The main aim of this research is to design a robust framework for image forensics that handles real-world complexities. The following goals have been set:

1. **Detection of Single-operation Manipulations:** Developing a basic detection model for identifying the difference between unaltered images and those manipulated by applying a specific operation, e.g., Median Filtering, by utilizing subtle pixel correlations which cannot be identified with an unaided eye.
2. **Tolerance of Post-compression Artifacts (JPEG):** As most digital photographs are compressed using JPEG, the proposed model must be trained to differentiate JPEG compression artifacts from the actual image manipulations.
3. **Reconstruction of a Manipulation Chain:** In real-life situations, many images have undergone several manipulations before publication. While detection has been limited to classifying images into tampered or genuine, the proposed project will try to reconstruct the exact chain of manipulations (e.g., Median Filtering → Compression → Blurring).
4. **Detection on Low-Resolution Images:** Forensics is not always performed on high-resolution images. This objective deals with detecting image manipulations on low-resolution or down-scaled images.

1.7 Organisation of the Thesis

Organization of the remainder of this thesis Chapter 2 lays down the theoretical and technical framework of image forgery forensics, namely, the concept of image manipulation forensics, the DCT transform and JPEG compression, residual learning and CBAM attention. Chapter 3 gives a detailed review and critique of the related literature, identifying the research gap that necessitates AFNet. Chapter 4 covers the technical details regarding the proposed method and the experimental setup. Chapter 5 provides an account of the experiments and their results, with both overall performance and class-wise performance. Chapter 6 discusses the results and their implications. Chapter 7 concludes the thesis.

CHAPTER 2

TECHNICAL BACKGROUND

This chapter provides the theory and technology that lies behind AFNet architecture. First, the nature of traces left by typical image manipulations is addressed, and the influence of the JPEG compression process on that is discussed. The next step is to outline the deep learning elements AFNet relies on, and to finish the chapter presents the dataset preparation.

2.1 Image Manipulation Operations and Forensic Fingerprints

Any image processing algorithm affects the intrinsic statistical behavior of the image in a unique way. Statistical patterns exhibited by natural images result from the camera's light-capturing process, its sensor response, texture distributions, and the interdependencies of neighboring pixels. Processing introduces deviations from those known statistical patterns, which become forensic fingerprints, recognizable traces of the manipulation. **Median Filtering (MF)** replaces each pixel by the median value among its neighbors, which makes it an efficient tool for impulse noise suppression. However, MF leads to the loss of the local pixel correlation behavior resulting in easily noticeable abnormalities in the noise residual power spectrum [5]. These abnormalities survive JPEG compression to some extent, making MF one of the easier to detect algorithms. **Gaussian Blur (GB)** performs convolution of the image with a Gaussian filter, gradually decreasing high-frequency content. The corresponding noise residual exhibits low-pass attenuation that is different from anomalies in the median filtering case and from JPEG blocking effects [6]. **Resampling (RS)**, involving scaling and rotation, modifies the sampling grid by means of interpolation. This causes interdependencies between pixels not present in the original image, resulting in the appearance of periodic peaks in the frequency domain [7]. **Additive White Gaussian Noise (AWGN)** superimposes zero mean Gaussian noise on the pixel values, thereby increasing the broadband noise level consistently across the entire signal. AWGN may also be added intentionally as an anti-forensic technique to obscure evidence of other alterations. **Contrast Enhancement and Histogram Equalization** alter the pixel intensities in

such a way as to distort the camera response function, leading to unique comb-shaped artifacts in the pixel intensity distribution histogram [8]. It should be noted that when transformations are performed one after another, their respective marks do not simply sum together. Each transformation changes the signal in a different manner, thus creating a complex signal fingerprint reflecting the type and order of all transformations applied. The results of applying median filtering followed by Gaussian blur will be distinctly different from applying Gaussian blur followed by median filtering, while both will be significantly different from just applying one filter or the other [2].

2.2 JPEG Compression and Its Forensic Impact

³⁴ JPEG is the most popular image encoding format used in consumer, industrial, and online imaging applications, and the importance of JPEG in forensic analysis arises in two respects: it adds its own exploitable artefacts, and it interacts with artefacts left behind by previous manipulations in a manner that conceals them. The JPEG technique involves partitioning an image into non-overlapping 8×8 pixel blocks, applying the Discrete Cosine Transform to the data in each block, quantizing the resulting coefficients according to a factor-quality based table, and finally, encoding them using entropy coding. The quantization step is inherently lossy each coefficient is rounded to the nearest grid value, and the rounding error is permanently discarded. This produces blocking discontinuities in the spatial domain and grid-like patterns in the DCT coefficient histograms, both of which remain statistically measurable even at high quality factors where they are invisible to the human eye. For chain detection, the critical challenge is that this quantization noise spectrally overlaps with the residual traces left by prior spatial-domain manipulations such as median filtering and Gaussian blurring. When the QFs are low, the masking effect could be so pronounced that it completely overwhelms the manipulation footprint [4, 10]. The problem becomes even more pronounced when multiple compression cycles occur at the acquisition stage and once more during online dissemination. This is the primary rationale behind approaches that model behavior in the DCT domain.

2.3 Deep Learning Building Blocks

AFNet draws on four interconnected deep learning concepts, each playing a specific role in the forensic detection pipeline.

2.3.1 Convolutional Neural Networks

CNNs represent an effective method for learning hierarchical, spatially-aware feature representations from data, and therefore naturally lend themselves to the forensic application due to the strong dependence of the discrimination signal on spatial information. Each layer learns filters to apply over its inputs to generate feature maps with increasing levels of abstraction the deeper into the network one goes, while the pooling process decreases the spatial resolution of these outputs and nonlinear activation functions allow for more advanced modelling. One of the more interesting forensically adapted CNNs is the constrained convolutional model developed by Boyar and Stamm [3]. This model involves the addition of a sum-to-zero normalization constraint on the weights of the first layer of the model during training. By doing so, the first layer of the model must immediately focus on detecting high-frequency residual information rather than low-frequency information that comprises the general smoothness of a given scene.

2.3.2 Residual Networks

The central challenge in using deep CNNs for forensics is that the forensic signal is much smaller in comparison to the major scene component. Given a conventional deep network that learns from raw pixel values, the network would tend to learn the scene semantics rather than forensic residuals since the scene component adds more to the loss function. In order to solve this problem, residual networks introduce a different architecture where each block learns the difference between the desired output and the input whereas a skip connection ensures that the input signal remains intact till it reaches the output [6].

2.3.3 DCT and Frequency-Domain Analysis

Discrete Cosine Transform serves as both the underlying mathematics of the JPEG encoding algorithm and a powerful forensic tool. Discrete Cosine Transform transforms every pixel block of size 8×8 to spatial frequency coefficients; where low-order coefficients contain information on the coarse structure of an image, and high-order coefficients describe textures and edges. One benefit of using the DCT transform is the creation of a pattern within the transformed coefficients even for operations that cannot be easily distinguished in the spatial domain [4]. CAS branch of AFNet directly uses DCT transform to extract frequency features from 25 DCT coefficient maps using dilated convolution.

2.3.4 Attention Mechanisms: CBAM

In regular convolutions, all channels and spatial locations have equal weightage. However, for forensic purposes, an equal weightage approach falls short since the discriminative features lie within specific frequency ranges, certain spatial locations such as edges and textures, and particular stages of processing. This problem is tackled by the Convolutional Block Attention Module (CBAM) [11], which uses channel attention followed by spatial attention. The channel attention mechanism makes use of global average and global max statistics on feature maps and employs a multi-layer perceptron to weigh these statistics; global average represents the general activation levels of channels, while the global max represents the maximum activations [13]. Following that, spatial attention locates the useful areas using channel-wise pooling, concatenation, and 7×7 convolutions with sigmoid activations. The large kernel size is suitable for accommodating the extended nature of forensic artifacts such as resampling artifacts. In the forensic scenario, channel attention filters out channels that have high contributions from JPEG blocking artifacts shared by all compressed images, while retaining channels that have manipulation-specific artifact information. Spatial attention, on the other hand, helps the model concentrate on edges and texture boundaries where those artifacts are likely to be found. There is evidence from previous forensic studies on splicing [11] and copy-move forgery [12] detections showing that this approach can help.

2.4 Dataset, Manipulation Categories, and Resolution Analysis

2.4.1 Dataset Construction

Both training, validation, and testing datasets are obtained from the RAISE ALL benchmark [14], which is publicly available and consists of 3,000 high-quality raw TIFF images under varying imaging conditions for diverse cameras. The choice of the RAISE ALL dataset is justified by the fact that using an uncompressed raw format guarantees a dataset completely void of any previous compression artifacts. First, the image undergoes a JPEG compression with quality factor $QF_1 = 90$ to mimic in-camera compression. Then, a particular manipulation is performed, followed by another JPEG compression with $QF_2 \in \{50, 60, 70, 80, 90\}$ to represent the image recompression when distributed. Finally, the image is transformed to grayscale and cut into non-overlapping patches of size 256×256 , resulting in 30,000 training samples, 8,000 validation samples, and 15,000 testing samples. For evaluating the model performance on cross-dataset evaluation, the models trained on RAISE ALL are tested on DRES-

DEN [15] and UCID [16].

2.4.2 Manipulation Categories

Six different manipulation types are investigated, all of which are forensic applications of image processing: **Median Filtering (MF)** is performed on kernels 3×3 and 5×5 . Apart from its usual application of suppressing noise in the image, median filtering is often utilized as an anti-forensic tool as it partially obfuscates the presence of other types of manipulations while preserving the characteristic artifact of its own [5]. **Gaussian Blur (GB)** is performed using kernels 3×3 and 5×5 . Gaussian blurring is frequently used for artistic effects, privacy protection, and simulating the shallow depth of field [6]. **Resampling (RS)** is done using bicubic interpolation at scale factors 1.2 and 1.5. This process is part and parcel of the editing process and also commonly used in image splicing when the areas need to be made compatible with each other in terms of scaling [7]. **Additive White Gaussian Noise (AWGN)** is added at standard deviations $\sigma \in \{0.5, 2.0\}$ to replicate the effects of acquisition noise and also intentional anti-forensic processing. **Unaltered** images have only the double JPEG compression procedure performed on them, acting as the negative control. There are two types of **compound chain**: $\text{MF} \rightarrow \text{GB}$ and $\text{GB} \rightarrow \text{MF}$, which form the basis of the major stress test cases. These chains combine two smoothing processes whose distinct fingerprints may partially overlap, and whose resultant effects on DCT energy are almost indistinguishable from a heavily compressed clean image.

2.4.3 Image Resolution and Forensic Implications

The level of resolution that forensic analysis is done impacts the efficacy of detection in several ways. The forensic fingerprint is localized and is a statistical anomaly within groups of pixels, noise, and DCT coefficients. The lower the resolution, the less number of pixels will be used to detect such anomalies. This work considers four patch resolutions: 32×32 , 64×64 , 128×128 , and 256×256 pixels. With the image size being 256×256 , the full spatial footprint of the manipulation artifacts is kept intact, and both channels are working to their full potential. With the image size being 128×128 , good detail is maintained but the noise pattern is starting to deteriorate. With the image size being 64×64 , the periodicity introduced by resizing and the blocking effect of JPEG artifacts cover fewer pixels and more emphasis will be placed on the frequency domain patterns and learned priors. The progressive spatial downsampling within both the NRAS and CAS streams from 256×256 through 128×128 down to 64×64 is a deliberate design choice that aggregates residual evidence across multiple spatial scales before classification, preserving discriminative information.

CHAPTER 3

RELATED WORK

The issue of identifying and analyzing the characteristics of manipulation in an image has received constant research focus for decades because of its significance in digital forensics and media authenticity purposes. The literature that is pertinent to the current research is categorized according to the following four themes: handcrafted methods in forensic techniques, use of deep learning techniques for image manipulation detection, manipulation chain detection algorithms, and attention models used in forensic techniques. A critical synthesis at the end of the chapter identifies the research gap that directly motivates AFNet.

3.1 Traditional Forensic Detection Methods

The earliest computational approaches to image manipulation detection were grounded in the design of handcrafted statistical features tailored to the known properties of specific manipulation operations. These methods were productive and yielded practically effective detectors for individual operations, but their inherent operation-specificity proved a fundamental barrier to generalisation.

In [5], Yuan et al. presented a blind forensic detection method based on spectral characteristics of the pixel-level autocorrelation function to analyse the effects of median filtering. Since the median filter leads to a flattening in the autocorrelation peak, this approach can detect the use of this filter even under moderate amounts of JPEG compression. Spectral analysis of median filtering was first established as a promising forensic solution, becoming a reference point for other residual-based techniques. Feng et al. [7] employed spectral energy density features based on the frequency spectrum as a means of detecting resampling. This method relied on the presence of periodicity in the form of peaks within the spectrum induced by the geometric interpolation used to achieve resampling. Cao et al. [8] provided features using histograms of pixels for contrast enhancement forensics. Their approach was based on observing the comb-shaped distribution of pixels after the application of such enhancements.

All of these approaches, when taken together, illustrated the inherent tractability of

single-operation forensic detection and delivered the characterizations needed by the later deep-learning techniques in order to be able to learn them automatically. However, their main drawback lay in the need to have a new detector created for each operation, which limited their ability to generalize.

3.2 Deep Learning Approaches to Manipulation Detection

The emergence of deep learning fundamentally transformed the forensic detection landscape by enabling the automatic learning of discriminative features directly from data, bypassing the need for hand-engineered feature descriptors and providing substantially improved generalisation across manipulation types and compression conditions.

Bayar and Stamm [3] introduced constrained convolutional neural networks in which the first convolutional layer was subject to a sum-to-zero normalisation constraint. This constraint forced the layer to suppress low-frequency scene content and produce output activations dominated by high-frequency residuals, precisely the domain in which manipulation signatures reside. The resulting networks demonstrated generalisable detection capability across a wide range of manipulation types without prior knowledge of the specific operation, because the constrained initialisation naturally aligned the network's inductive biases with the forensic task. Boroumand and Fridrich [6] subsequently demonstrated that deep residual architectures adapted to the forensic setting achieved strong performance by concentrating network capacity on manipulation-induced noise residuals rather than scene semantics, confirming the value of residual learning for forensic feature extraction.

Rana et al. [17] introduced the Multi-Scale Residual Dense Network (MSRD-CNN), a more architecturally sophisticated single-stream framework comprising three sequential components: a multi-scale residual preprocessing module for adaptive noise extraction, a hierarchical feature extraction body employing Feature Extraction Blocks with dense residual connections, and a softmax classification head. MSRD-CNN achieved 97.07% and 97.48% accuracy on the BOSSBase and DRESDEN datasets for general-purpose manipulation detection, establishing that multi-scale residual aggregation in the spatial domain could yield excellent broad manipulation classification. However, as its per-class performance analysis reveals, the absence of DCT-domain modelling causes sharp degradation on compound smoothing chains under heavy JPEG compression, underscoring that spatial-only architectures are insufficient for the forensic chain detection problem.

3.3 Manipulation Chain Detection

Detecting the particular case of manipulative sequences, meaning more than just the fact that the image was manipulated, but determining which type of operations and in what sequence they were used, is an area of significant research interest because of its importance in practice.

Chu et al. [2] provided a foundational theoretical and empirical analysis establishing that different orderings of the same operations produce measurably distinct forensic fingerprints. This work formalized the theoretical basis for order-aware chain detection and demonstrated that classifiers insensitive to operation ordering will systematically fail in multi-operation scenarios, motivating the development of architectures that explicitly model the sequential structure of manipulation chains.

Liao et al. [4] developed a two-stream CNN for robust detection of operator chains under realistic compression conditions, demonstrating that joint spatial-domain and DCT-domain processing significantly outperformed single-stream approaches for JPEG-compressed inputs. Yu et al. [10] proposed a two-stream network specifically for JPEG manipulation classification, showing that parallel processing of image-domain and DCT-domain representations achieved high chain classification accuracy by capturing the complementary information available in both domains.

Kadha et al. [9] synthesised these findings into the ReMReNet dual-stream architecture, combining a Noise Residual Extraction stream for spatial-domain forensic traces with a Compression Feature Extraction stream for DCT-domain analysis. This design achieved above 91.98% accuracy across various manipulation chain categories, establishing the dual-stream paradigm as the leading approach. However, in ReMReNet, all extracted features' channels and spatial locations are implicitly weighted equally in intra-stream processing and inter-stream fusion, limiting its ability to discriminate in forensics situations in which the important information is only present in certain channels and locations.

3.4 Attention Mechanisms in Image Forensics

It has been found that attentional systems can be successfully employed in many computer vision problems, as well as in forensics applications, showing how the capacity to assign weights to informative attributes readily extends to the domain of forensics.

SE-net [13], on the other hand, pioneered the technique of channel-wise recalibration based on global pooling and gating operation that learns scalar weights representing the importance of the channel features given their global properties. When

considering forensic applications, channel attention of SE-type facilitates the filtering out of channels carrying scene information irrelevant to the forensic task while simultaneously enhancing channels capturing relevant forensic data. In addition, the concept of CBAM [1] extends channel attention by including spatial attention into it, thereby offering more holistic feature refinement. Multiple vision tasks confirmed experimentally that CBAM consistently exceeds SE-channel attention performance, encouraging its incorporation into AFNet.

In their study, Zhou et al. [11] have proven that attention-based localization was able to enhance the detection of image splicing, as the attention mechanism enabled the representational capability to be focused on the boundaries of the real and tampered image regions. The effectiveness of attention-based localization was further verified for copy-move forgery using different images and compressions [12].

3.5 Research Gap and Motivation

A careful examination of the literature reveals a well-defined gap at the intersection of dual-stream compression-aware forensics and adaptive attention-based feature weighting. The most effective existing chain detection architectures [4, 9, 10] correctly adopt the dual-stream paradigm but apply uniform implicit weighting to all extracted features. Simultaneously, the demonstrated effectiveness of CBAM-style attention in related forensic tasks [11, 12] suggests that similar gains are achievable for chain detection, yet this combination has not been systematically studied. AFNet directly fills this gap by integrating CBAM at multiple complementary locations within a dual-stream architecture, combining the demonstrated strengths of both design strategies in a single coherent framework. Table 3.1 summarises the principal methodological streams and positions AFNet within the landscape.

The table makes clear that AFNet uniquely combines dual-stream compression-aware architecture with systematic multi-point attention integration, occupying a design point that has not previously been explored in the published literature.

Table 3.1. Comparison of Major Image Forensic Detection Approaches.

Representative Work	Key Strengths	Limitations	Method Category
Yuan et al. [5], Feng et al. [7]	Interpretable and tailored to specific operations	Cannot generalise across different manipulation types	Handcrafted Statistical Features
Bayar & Stamm [3]	Suppresses scene content, works across manipulation types	Single-stream with no frequency-domain modelling	Constrained CNNs
Rana et al. [17]	Rich multi-scale spatial feature extraction	Single-stream, degrades under heavy JPEG compression	Multi-Scale Residual Networks
Liao et al. [4], Yu et al. [10]	Combines spatial and frequency-domain representations	All extracted features are weighted equally	Two-Stream Networks
Kadha et al. [9]	Best chain detection accuracy before this work	Uniform feature weighting applied at the fusion stage	Dual-Stream with DCT
Zhou et al. [11], Zhong & Pun [12]	Spatially selective, focuses on tampered regions	Not extended to manipulation chain detection	Attention-Based Forensics
Our work	Dual-stream with multi-point attention at every processing stage	Marginal computational overhead only	AFNet (Proposed)

CHAPTER 4

PROPOSED METHODOLOGY

³⁹ This chapter provides a comprehensive technical description of the AFNet architecture and the experimental pipeline in which it is evaluated. The presentation proceeds from the dual-stream backbone through the CBAM module and its three integration points, the complete network assembly with classification head, the training strategy, and the dataset construction and evaluation protocol.

4.1 Overview of the AFNet Architecture

AFNet processes 256×256 grayscale image patches through two parallel streams that extract complementary forms of forensic evidence before combining them for classification. The Noise Residual Attention Stream (NRAS) operates on the raw pixel representation to extract spatial-domain forensic fingerprints through hierarchical residual learning. The Compression Artifact Stream (CAS) operates in the DCT frequency domain to model the characteristic patterns introduced by JPEG compression and their interaction with prior manipulation traces.³⁸ The outputs of both streams are fused, refined by a cross-stream attention module, and passed through a classification head to produce the final manipulation chain prediction.³ The overall architecture of AFNet is illustrated in Figure 4.1.

4.2 Noise Residual Attention Stream

²¹ The NRAS receives a single-channel grayscale input of shape $1 \times 256 \times 256$. Two initial 3×3 convolutional layers with 64 output channels, each followed by Batch Normalisation and ReLU nonlinearity, perform the first-stage feature extraction, establishing a residual-enriched representation that de-emphasises smooth low-frequency scene content in favour of the high-frequency statistical anomalies introduced by manipulation operations.

The main body of the NRAS is organised into four sequential processing stages. Stages 1 and 2 each consist of four standard residual blocks operating at 32 channels

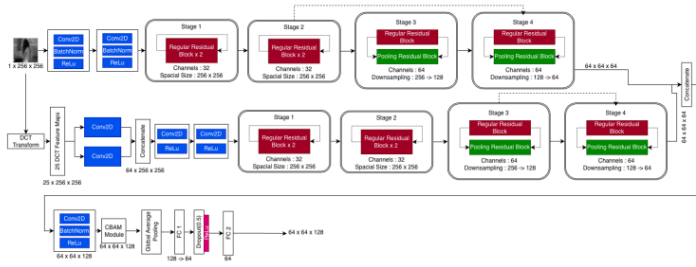


Figure 4.1. AFNet dual-stream architecture with NRAS and CAS streams, DCT transformation, CBAM-enhanced fusion, and classification head.

while maintaining the full 256×256 spatial resolution, enabling the extraction of fine-grained noise textures at the original scale. Each residual block comprises two 3×3 convolutional layers with Batch Normalisation and ReLU, connected by a skip connection that carries the block input to the output. Stages 3 and 4 each begin with a pooling residual block that progressively downsamples spatial dimensions from $256 \rightarrow 128$ and $128 \rightarrow 64$ pixels respectively while increasing the channel count to 64. The output tensor of the NRAS is thus of shape $64 \times 64 \times 64$. The stream architecture is illustrated in Figure 4.2.

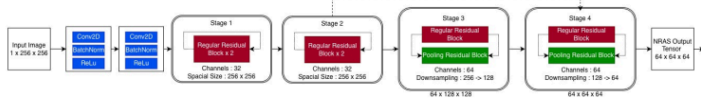


Figure 4.2. NRAS stream with convolutional layers, regular residual blocks (cyan), and pooling residual blocks (orange).

4.3 Compression Artifact Stream

The CAS operates in the DCT frequency domain to explicitly model the interaction between JPEG compression artefacts and prior spatial-domain manipulation traces. The stream begins with an 8×8 block-based DCT transformation of the grayscale input, which generates 25 frequency feature maps of shape $25 \times 256 \times 256$. Each of these maps corresponds to a specific DCT coefficient position across all 8×8 blocks of the image, providing a representation in which the spatial layout of frequency-domain information is directly accessible to subsequent convolutional processing.

These 25 frequency maps are processed through two parallel convolutional branches, one with a standard dilation rate of 1 and one with a dilation rate of 4, each producing 32 output channels. The standard-dilation branch captures local frequency patterns within and immediately around individual DCT blocks. The dilated branch, with its expanded receptive field, is specifically designed to detect the extended periodic compression artefacts that arise across multiple adjacent DCT blocks and that standard convolutions with small kernels cannot identify. Concatenating the two branch outputs along the channel dimension yields a $64 \times 256 \times 256$ representation that benefits from both local and extended-range frequency analysis. Two 3×3 convolutional layers followed by four residual stages matching the NRAS structure reduce the spatial dimensions to produce a $64 \times 64 \times 64$ output, dimension-matched to the NRAS output for subsequent fusion. The CAS stream architecture is illustrated in Figure 4.3.



Figure 4.3. CAS stream with DCT transformation, parallel dilated convolutions, and residual blocks.

4.4 CBAM Module

The CBAM module [1] applied at each integration point in AFNet performs sequential channel and spatial attention refinement. Given an input feature tensor $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$, the channel attention sub-module computes a C -dimensional weighting vector using both global average pooling and global max pooling to capture complementary statistics of each channel, processing both through a shared MLP with bottleneck ratio $r = 16$ and summing with a sigmoid:

$$\mathbf{M}_c(\mathbf{F}) = \sigma(\text{MLP}(\text{AvgPool}(\mathbf{F})) + \text{MLP}(\text{MaxPool}(\mathbf{F}))) \quad (4.1)$$

The resulting channel-refined feature $\mathbf{F}' = \mathbf{M}_c(\mathbf{F}) \otimes \mathbf{F}$ is passed to the spatial attention sub-module, which computes a $H \times W$ spatial weighting map by channel-wise pooling, concatenation, and 7×7 convolution:

$$\mathbf{M}_s(\mathbf{F}') = \sigma(\text{Conv}_{7 \times 7}([\text{AvgPool}_c(\mathbf{F}'), \text{MaxPool}_c(\mathbf{F}')])) \quad (4.2)$$

The final output is:

$$\mathbf{F}'' = \mathbf{M}_s(\mathbf{F}') \otimes \mathbf{F}' \quad (4.3)$$

This two-stage refinement simultaneously suppresses uninformative channels and uninformative spatial positions. In the forensic context, channel attention suppresses channels dominated by JPEG blocking artefacts common to all compressed images (and therefore non-discriminative for chain classification) while amplifying channels encoding the specific statistical anomalies of each manipulation type. Spatial attention then focuses on the image regions, the edges, texture boundaries, and periodic patterns, where these anomalies are most spatially concentrated. Figures 4.4 and 4.5 illustrate the CBAM channel attention and spatial attention module architectures respectively.

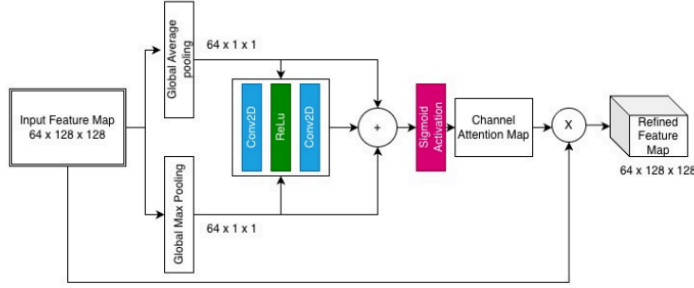


Figure 4.4. CBAM Channel Attention Module: uses global average pooling and global max pooling, processed through a shared MLP with bottleneck ratio $r = 16$, and sigmoid activation to produce the channel attention map [1].

4.5 Multi-Point CBAM Integration Strategy

The architecturally distinctive feature of AFNet is the integration of CBAM at three complementary and independently motivated locations, each addressing a specific representational need in the overall forensic pipeline.

4.5.1 CBAM within Residual Blocks

The first and most impactful integration point inserts a CBAM module within every residual block, positioned between the output of the second convolutional layer and the residual skip addition. By refining the transformed residual before it is added to the skip-connection path, the attention module prevents uninformative or noise-dominated residuals from accumulating as the signal propagates through successive blocks. Ablation experiments confirm this placement provides the largest individual contribution to overall accuracy, consistent with the interpretation that per-block refinement yields

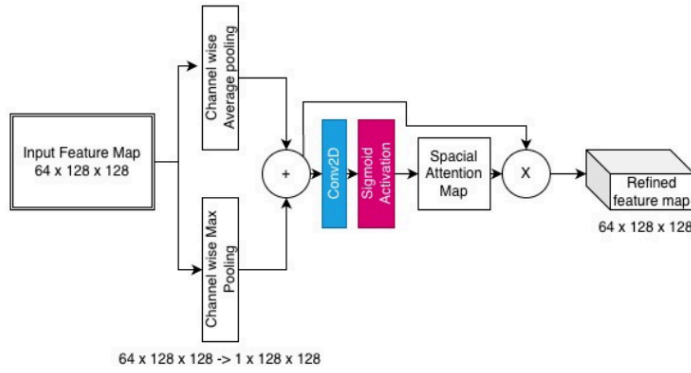


Figure 4.5. CBAM Spatial Attention Module: uses channel-wise average pooling and max pooling, concatenation, 7×7 convolution, and sigmoid activation to produce the spatial attention map [1].

compounding benefits across the full network depth. Figures 4.6 and ?? illustrate the CBAM-augmented regular and pooling residual blocks respectively.

4.5.2 CBAM after Initial Convolutional Layers

The second integration point places CBAM immediately after the initial convolutional layers of each stream, before features enter the main residual body. At this early stage, the module acts as an adaptive gating mechanism that evaluates the initial extracted features and suppresses channels and regions dominated by scene content or camera-system noise, directing the subsequent residual processing toward the portions of the feature representation that carry forensic information. This early-stage filtering allows all subsequent residual blocks to concentrate on refining already-selected forensic features rather than sifting through raw unfiltered activations.

4.5.3 CBAM within the Cross-Stream Fusion Layer

The third integration point applies CBAM to the fused feature tensor produced by concatenating the NRAS and CAS outputs and passing them through a 1×1 convolutional projection. The NRAS and CAS streams carry qualitatively different types of forensic evidence, spatial noise residuals and DCT compression patterns respectively and their relative discriminative informativeness varies with the specific manipulation chain and the JPEG quality factor. The cross-stream fusion CBAM learns to reweight the com-

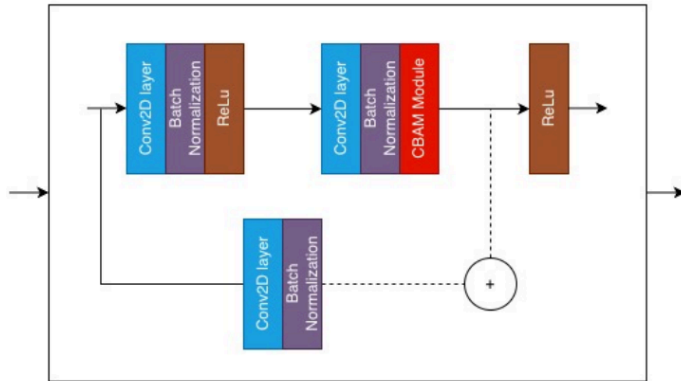


Figure 4.6. Regular residual block with CBAM (magenta) after the second convolution and before the skip addition.

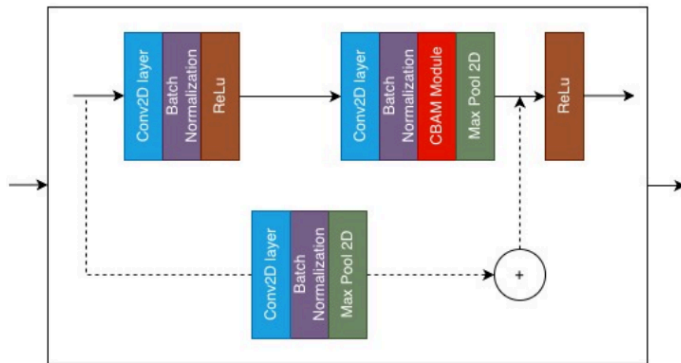


Figure 4.7. Pooling residual block with CBAM and max pooling (red) on the skip path for spatial downsampling.

binned 128-channel representation according to the input-specific evidence, amplifying NRAS channels when spatial traces are well-preserved and CAS channels when frequency-domain evidence is more discriminative. This adaptive cross-stream balance is a capability that fixed-weight fusion architectures fundamentally cannot provide.

4.6 Complete Architecture and Classification Head

After the cross-stream CBAM module, the attention-refined $128 \times 64 \times 64$ feature tensor is passed through global average pooling, collapsing the spatial dimensions to produce a 128-dimensional feature vector. This vector is processed by a fully connected classification head with the structure $128 \rightarrow 64 \rightarrow \text{num_classes}$, with a Dropout layer at rate 0.5 between the two linear layers to regularise training. The complete architectural assembly is illustrated schematically in Figure 4.1.

4.7 Training Strategy

AFNet is trained using cross-entropy loss with label smoothing at $\epsilon = 0.1$. Label smoothing prevents the network from becoming overconfident on the confusable compound chain classes particularly MF $\rightarrow$ GB and GB $\rightarrow$ MF, which share many statistical properties by replacing the hard one-hot target with a softened distribution that assigns a small probability mass to all classes. The optimiser is AdamW with base learning rate 0.0001 and weight decay 0.0005, providing adaptive per-parameter learning rates with L2 regularisation through the weight decay term. A step learning rate scheduler reduces the learning rate by 50% every 10 epochs, enabling large initial parameter updates followed by progressively finer refinements as training converges.

The effective training batch size is 16, achieved through gradient accumulation where necessary. Data augmentation during training includes horizontal flipping and random rotation within $\pm 10^\circ$, increasing training distribution diversity without distorting the manipulation-specific forensic features that the network is trained to detect. Training is halted by early stopping with patience of 10 epochs, preventing overfitting to the training partition. All experiments are conducted on an NVIDIA RTX 4080 using PyTorch 2.0 with mixed-precision training.

4.8 Dataset Construction and Experimental Setup

The first evaluation set comprises images extracted from the RAISE ALL benchmark [14] that includes 3,000 high-quality raw TIFF images. The images in this dataset go

through the following processing steps: the initial JPEG compression using the quality factor of $QF_1 = 90$; applying the required manipulation; secondary JPEG compression using $QF_2 \in \{50, 60, 70, 80, 90\}$; converting the images to grayscale; and extracting non-overlapping 256×256 patches. These processing steps mimic actual forensic settings where images are compressed upon acquisition, manipulated, and subsequently re-compressed before being disseminated.

The study encompasses six manipulation categories: Median Filtering (kernel sizes 3×3 and 5×5), Gaussian Blur (kernel sizes 3×3 and 5×5), Resampling (bicubic at scale factors 1.2 and 1.5), Additive White Gaussian Noise (AWGN) with $\sigma \in \{0.5, 2.0\}$, and Unaltered. The compound chain categories MF $\rightarrow$ GB and GB $\rightarrow$ MF are included to specifically evaluate chain detection performance. The dataset is partitioned into 30,000 training, 8,000 validation, and 15,000 test images. The DRESDEN [15] and UCID [16] datasets serve as held-out benchmarks for cross-dataset generalisation evaluation.

All baseline methods, Bayar and Stamm [3], Yu et al. [10], Liao et al. [4], MSRD-CNN [17], and Kadha et al. [9] are re-evaluated on RAISE ALL under the same training and compression settings to ensure fair comparison. For MSRD-CNN, results are reproduced from RAISE ALL using the published architecture and the same patch extraction pipeline, since the original results were reported on BOSSBase and DRESDEN.

EXPERIMENTAL RESULTS AND EVALUATION

This chapter presents a comprehensive account of the experimental results obtained by applying AFNet and the evaluated baseline systems to the RAISE ALL benchmark and the cross-dataset evaluation corpora. The presentation covers overall accuracy comparison, per-class performance analysis, ablation studies that isolate individual CBAM contributions and compare attention formulations, cross-dataset generalisation evaluation, and computational overhead analysis.

5.1 Overall Performance on RAISE ALL

Table 5.1 presents the training, validation, and test accuracy results for all evaluated methods on the RAISE ALL benchmark, aggregated across all five compression quality factor settings.

Table 5.1. Overall Accuracy Comparison on RAISE ALL (%).

Model	Train	Val	Test
Bayar & Stamm [3]	91.20	90.85	91.14
Yu et al. [10]	93.45	93.10	93.62
Liao et al. [4]	94.30	93.80	94.21
MSRD-CNN [17]	95.10	94.72	95.33
Kadha et al. [9]	96.24	96.15	96.56
AFNet (Ours)	97.75	98.56	97.17

AFNet achieves a test accuracy of 97.17%, surpassing all evaluated baseline methods by a clear margin. The overall performance progression from Bayar and Stamm [3] through MSRD-CNN [17] and Kadha et al. [9] to AFNet reflects a coherent line of architectural development, with each step addressing the limitations of the previous design. The constrained CNN of Bayar and Stamm achieves 91.14%, demonstrating the value of content-suppressive feature learning but limited by its single-stream, spatial-

only design. MSRD-CNN extends this to 95.33% through multi-scale residual aggregation but remains restricted to the spatial domain. Kadha et al. adds DCT-domain modeling through a dual-stream design, reaching 96.56%. AFNet’s further improvement to 97.17% through adaptive feature weighting confirms that attention-based feature selection provides significant additional benefit beyond what dual-stream design alone achieves.

It is noteworthy that AFNet’s validation accuracy of 98.56% exceeds its test accuracy of 97.17%, suggesting that the CBAM-induced regularisation which suppresses uninformative features and directs representational capacity toward forensically relevant signals reduces overfitting relative to architectures without attention. The 1.39 percentage point gap between validation and test performance is consistent with normal generalisation behaviour and does not indicate over-tuning to the validation set.

5.2 Per-Class Performance Analysis

Table 5.2 presents the per-class accuracy, precision, recall, and F1-score for MSRD-CNN [17], Kadha et al. [9], and AFNet across all five manipulation categories on the RAISE ALL test set.

Table 5.2. Per-Class Performance Comparison on RAISE ALL (%).

Class	MSRD-CNN [17]				Kadha et al. [9]				AFNet (Ours)			
	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
Original	96.10	95.80	96.10	95.95	98.86	98.45	98.86	98.65	99.60	99.16	99.60	99.38
MF	95.42	95.10	95.42	95.26	97.64	97.24	97.64	97.44	98.47	98.30	98.47	98.39
GB	94.85	94.50	94.85	94.67	96.48	96.15	96.48	96.31	97.85	97.55	97.85	97.70
MF→GB	91.23	88.60	91.23	89.90	95.96	92.54	95.96	94.22	97.79	94.82	97.79	96.28
GB→MF	84.10	88.20	84.10	86.10	89.56	93.45	89.56	91.46	92.15	96.04	92.15	94.05
Average	92.34	92.44	92.34	92.38	95.70	95.57	95.70	95.62	97.17	97.17	97.17	97.16

The per-class results reveal an interpretable and theoretically coherent pattern across all three systems. The single-operation categories (MF, GB) and the unaltered class are handled reliably by all three architectures, reflecting the relative simplicity of detecting well-characterised, isolated forensic fingerprints even in the presence of JPEG compression. MSRD-CNN achieves strong results on these categories, with 95.42% for MF and 94.85% for GB, consistent with the effectiveness of multi-scale spatial residual processing for single-operation scenarios.

The compound manipulation chains reveal the most diagnostically significant architectural differences. MSRD-CNN’s performance degrades sharply to 91.23% on

MF→GB and further to 84.10% on GB→MF, a decline of 11.32 percentage points from its best single-class result. This degradation reflects the inability of a single-stream spatial-domain architecture to model the complex interaction between compound smoothing traces and heavy JPEG compression artefacts. The dual-stream design of Kadha et al. substantially improves chain class performance through DCT-domain modeling, reaching 95.96% on MF→GB and 89.56% on GB→MF. However, the 6.40 percentage point gap between its best and worst chain classes indicates that equal-weight feature fusion still leaves significant discriminative capacity unrealised.

AFNet’s CBAM-augmented attention architecture closes these gaps substantially. The 97.79% accuracy on MF→GB represents improvements of 1.83 percentage points over Kadha et al. and 6.56 points over MSRD-CNN. The 92.15% accuracy on GB→MF represents improvements of 2.59 points over Kadha et al. and 8.05 points over MSRD-CNN. The 99.60% accuracy and 99.16% precision on unaltered images further demonstrate that the attention mechanism effectively suppresses false-positive detections on clean images, a critical property for forensic systems operating under real-world conditions where false alarms carry material costs.

5.3 Ablation Studies

Table 5.3²⁰ presents the results of the ablation study conducted to isolate the individual and combined contributions of the three CBAM integration points and to compare CBAM against alternative attention formulations.

The ablation results provide several important and consistent findings. Every individual CBAM placement yields a positive contribution over the no-attention baseline, confirming that each integration point provides genuine and non-redundant forensic benefit. CBAM within the residual blocks provides the largest individual contribution, improving test accuracy by 0.39 percentage points over the baseline. This result is consistent with the interpretation that per-block feature refinement produces compounding benefits as the refined representations propagate through successive network layers, preventing uninformative residuals from accumulating. CBAM after fusion provides the smallest individual contribution (0.22 points), reflecting that the most discriminative features have already been selectively amplified by the within-stream attention modules by the time the fusion stage is reached.

Critically, the full model combining all three placements achieves 97.17%, which exceeds the simple additive sum of the individual incremental contributions, demonstrating genuine synergistic interaction. The early-stage gating attention allows the residual block attention to operate on already-refined feature maps, and the cross-

Table 5.3. Ablation Study: Effect of Attention Placement and Formulation on Validation and Test Accuracy (%).

Configuration	Val	Test
<i>(a) Attention Placement Ablation</i>		
Baseline dual-stream network with no attention modules applied at any stage	96.15	96.56
CBAM inserted inside every residual block of both NRAS and CAS streams, with no attention at the fusion or early-stage layers	97.24	96.95
CBAM applied only at the cross-stream fusion layer where NRAS and CAS features are concatenated, with no attention inside the streams	96.85	96.78
CBAM placed after the initial convolutional layers at the entry point of both NRAS and CAS streams, with no attention in residual blocks or at fusion	97.56	97.04
Full AFNet model with CBAM at all three locations: inside every residual block, after the early-stage convolutional layers of both streams, and at the cross-stream fusion layer	98.56	97.17
<i>(b) Attention Formulation Ablation</i>		
Squeeze-and-Excitation (SE) modules [13] applied at all three positions; SE performs channel-only recalibration using global average pooling followed by a bottleneck MLP, without any spatial attention component	97.64	96.82
Channel attention component of CBAM applied alone at all three positions; uses both global average and global max pooling with a shared MLP to reweight feature channels, but omits the spatial attention step	97.38	96.91
Spatial attention component of CBAM applied alone at all three positions; uses channel-wise average and max pooling followed by a 7×7 convolution to highlight informative spatial regions, but omits the channel attention step	97.12	96.74

stream fusion attention benefits from the more discriminative NRAS and CAS representations produced by the preceding two attention stages. This synergy is the key architectural rationale for the multi-point integration strategy.

The comparison with alternative attention formulations provides further validation. SE modules achieve 96.82%, channel-only CBAM variants achieve 96.91%, and spatial-only variants achieve 96.74%, all below the full CBAM result of 97.17%. The ordering spatial-only < SE < channel-only < CBAM confirms that joint channel-spatial attention provides complementary benefits that neither type alone can fully deliver. Spatial attention is uniquely important for localising the spatially concentrated forensic patterns of operations such as median filtering and resampling, while channel attention is uniquely important for suppressing channels dominated by JPEG noise that is common to all compressed images regardless of manipulation.

5.4 Cross-Dataset Generalisation

Table 5.4 presents the accuracy results on the DRESDEN and UCID benchmarks for models trained exclusively on RAISE ALL.

Table 5.4. Cross-Dataset Evaluation: Accuracy on DRESDEN and UCID (%).

Model	DRESDEN [15]	UCID [16]
MSRD-CNN [17]	94.20	93.10
Kadha et al. [9]	94.56	93.84
AFNet (Ours)	95.84	95.12

AFNet achieves 95.84% on DRESDEN and 95.12% on UCID, with consistent margins of approximately 1.28 percentage points over Kadha et al. on both benchmarks. These cross-dataset results are important for two reasons. First, they confirm that the CBAM-learned attention patterns generalise across different camera systems and imaging conditions, providing strong evidence that the attention mechanism amplifies genuinely forensic, domain-invariant features rather than RAISE ALL-specific statistical regularities. Second, the consistency of the AFNet advantage across two independent external benchmarks rules out the possibility that the RAISE ALL results reflect overfitting to the specific characteristics of the training dataset. The ability to generalise across datasets is a prerequisite for practical forensic deployment, where the camera device and acquisition conditions of the evidence image will in general differ from those of the training corpus.

5.5 Computational Overhead Analysis

The CBAM augmentation introduces a total of 0.08M additional parameters across all three integration points, representing a 1.0% increase in parameter count over the baseline dual-stream architecture. Training time is extended by 3.6%, and inference latency increases by 1 millisecond per batch, from 45ms to 46ms (a 2.2% relative increase). Given that AFNet achieves a 0.61 percentage point improvement over the strongest prior architecture at this marginal computational cost, the attention augmentation provides an exceptionally favourable performance-to-overhead trade-off and is fully viable for practical forensic deployment. The minimal parameter overhead also reduces the risk of overfitting, which is consistent with the regularisation effect observed in the validation performance comparison.

DISCUSSION AND ANALYSIS

This chapter critically interprets the experimental findings of Chapter 5 in their broader technical, methodological, and practical contexts. The discussion addresses the architectural factors explaining AFNet’s performance, the mechanisms by which attention addresses the compound chain detection challenge, the implications of the cross-dataset results, and the limitations of the present study.

6.1 Architectural Development and the Role of Each Design Decision

The hierarchical performance witnessed among the reviewed methods follows a well-defined architectural evolution, which is theoretically understandable. First, Bayar and Stamm [3] showed that the addition of content-suppressive convolutional constraints greatly helped in manipulation detection due to the matching inductive biases and residual forensics in the image. Second, with MSRD-CNN [17], the single stream paradigm was enhanced through multi-scale residual aggregation such that hierarchies of richer spatial representations improved overall accuracy in detecting generalized manipulations. Third, Kadha et al. [9] made the fundamental shift from single to double streams, understanding the need to explicitly model the DCT domain because, through JPEG compression, frequencies arise that cannot be separated from manipulated images within a spatial stream only. Finally, AFNet takes the architecture forward by incorporating content-aware attention mechanisms to guide representation at different levels of processing.

From the results of the ablation analysis, we can see that no component on its own will work effectively for AFNet. The complete function of AFNet arises from a combination of the synergy between dual streams processing, DCT domain modeling, and multi-point attention fusion. The lack of any of these elements will affect the performance adversely, but the adverse effects do not simply combine, because the cross-stream fusion attention works better with intra-stream attention representation and the intra-stream residual block attention works best after the gating attention process.

6.2 Channel and Spatial Attention in the Forensic Context

The different functions of channel attention and spatial attention deserve individual consideration. With the use of channel attention, the network can dampen the feature channels that are affected by the JPEG compression artifacts, which are present in any image due to compression and have no discriminative information about the manipulation chains. The reduction in importance of such channels will help improve the signal-to-noise ratio for the manipulation-based characteristics, which remain in other feature channels. It is especially relevant to consider channel attention when working with the CAS stream, as many of the DCT maps contain only quantization noise patterns.

This is augmented by the use of spatial attention that detects and enhances areas of the image that have high spatial concentration of manipulation artifacts. Median filter and Gaussian blur produce artifacts mainly along edge and texture edges where the changes are maximum. Resampling artifacts are spread out throughout the image but still retain a specific spatial periodical nature. The 7×7 spatial attention mask is sufficient to detect such patterns, making it possible to localize the parts of the image most useful for classification.

The attention mechanism across the streams in the fusion stage brings about yet another benefit specific to the dual-stream framework, that of adjusting the importance weightage of NRAS and CAS features based on the extent of manipulation and compression of each particular image fed into the system. It is the very thing that a fixed weight fusion architecture is unable to achieve and contributes greatly to the superior performance of AFNet.

6.3 The Compound Smoothing Chain Detection Problem

However, since the GB→MF pair poses a consistently difficult problem, a special examination is required here. Indeed, the methods of Gaussian blurring and median filtering have an effect of reducing image frequencies, but by using them consecutively, one can achieve a combined effect that makes the energy distribution in the DCT domain similar to that of a highly compressed original image. This resemblance is the root cause of the systematic difficulty: the network must distinguish between a clean but heavily compressed image and an image that has undergone compound smoothing before compression, based on statistical differences that are progressively attenuated at lower quality factors.

The progression of GB→MF accuracy from 84.10% (MSRD-CNN) through 89.56%

(Kadha et al.) to 92.15% (AFNet) reflects the compounding benefits of each architectural addition. The absence of DCT-domain modeling in MSRD-CNN accounts for the largest single gap; adding dual-stream frequency analysis in Kadha et al. provides a substantial improvement; and the additional adaptive weighting in AFNet narrows the gap further. The 7.83 percentage point improvement of AFNet over MSRD-CNN on this single class represents the most significant per-class gain achieved across all evaluated architecture comparisons, and it is directly attributable to the combination of DCT-domain modeling and attention-based feature amplification.

6.4 Generalisation and Domain Invariance

The consistent performance advantage of AFNet over all competing architectures on DRESDEN and UCID provides strong evidence that the CBAM-learned attention weights encode domain-invariant forensic features rather than RAISE ALL-specific statistical regularities. The approximately 1.28 percentage point advantage over Kadha et al. on both external datasets, a smaller margin than on RAISE ALL itself but still consistent and statistically meaningful suggests that the attention mechanism acts as a regulariser, reducing the network’s tendency to exploit dataset-specific cues that do not transfer across camera systems and imaging conditions.

This generalisation property is a practical prerequisite for deployment in real-world forensic workflows, where the camera provenance of the evidence image is rarely controlled and may differ substantially from the training corpus. The cross-dataset evaluation provides the methodological evidence needed to support AFNet’s credibility as a forensically practical system rather than merely a benchmark-optimised model.

6.5 Limitations

Several limitations of this study should be acknowledged for completeness and to motivate future research. The evaluation is restricted to grayscale images; the extension to colour imaging, which provides additional forensic information in the chrominance channels, remains unexplored. The compound smoothing chain $GB \rightarrow MF$ achieves 92.15% accuracy, indicating that the problem is not fully solved and that meaningful headroom for improvement remains. The evaluation is limited to two-operation chains; the behaviour of AFNet on longer chains of three or more operations is unknown and merits systematic investigation. Finally, the robustness of AFNet against adversarial anti-forensic post-processing has not been evaluated, and such evaluation is important for understanding the security properties of the proposed system.

11 CHAPTER 7

CONCLUSION AND FUTURE WORK

7.1 Conclusion

This thesis has presented AFNet, an attention-driven dual-stream convolutional network for detecting image manipulation chains in JPEG-compressed environments. The central contribution is the systematic integration of the Convolutional Block Attention Module at three distinct and complementary locations within a dual-stream forensic architecture: within every residual block of both the Noise Residual Attention Stream and the Compression Artifact Stream, after the initial convolutional layers of each stream, and within the cross-stream feature fusion layer. This multi-point attention strategy enables hierarchical, content-adaptive feature selection across the full depth of the processing pipeline, simultaneously suppressing compression-induced noise, amplifying subtle inter-chain discriminative features, and dynamically balancing spatial-domain and frequency-domain evidence according to the specific characteristics of each input.

Comprehensive evaluation on the RAISE ALL benchmark demonstrates that AFNet achieves 97.17% overall test accuracy, establishing a new state-of-the-art result that surpasses all evaluated baselines. The performance gains are most pronounced on the compound smoothing chains that most severely challenge prior architectures: 97.79% accuracy on MF→GB chains and 92.15% on GB→MF chains, compared to 84.10% achieved by the strongest single-stream competitor and 89.56% by the strongest dual-stream baseline without attention. Cross-dataset evaluation on DRESDEN and UCID confirms generalisation across different camera systems and imaging conditions, with consistent margins of approximately 1.28 percentage points over the strongest prior approach. Ablation studies establish that all three CBAM integration points contribute non-redundantly and that their combination yields synergistic benefits exceeding the sum of individual contributions. The complete attention augmentation introduces only 0.08M additional parameters and 2.2% additional inference latency, confirming that the attention-based enhancement is computationally negligible relative to its forensic benefit.

Taken together, these results establish that the combination of dual-stream compression-aware architecture with systematic multi-point CBAM integration is a highly effective design strategy for forensic manipulation chain detection, and that adaptive feature weighting provides significant and consistent performance improvements over uniform-weight dual-stream baselines. The proposed approach offers direct practical benefits for verifying social media content, authenticating legal and journalistic evidence, and moderating images in operational forensic pipelines.

7.2 Future Work

Several directions for future research emerge from this thesis. Extending AFNet to process colour images would enable exploitation of additional forensic information available in the chrominance channels, particularly relevant for contrast enhancement and saturation-based manipulations. Exploring multi-scale attention mechanisms that operate across multiple DCT coefficient groups simultaneously could capture richer frequency-domain interactions and may further close the performance gap on the challenging GB→MF category. Investigating adversarial training against anti-forensic post-processing would quantify and improve the robustness of AFNet in adversarial deployment settings. Extending the evaluation framework to chains of three or more operations would address a practical gap, since real-world forgeries may involve more complex processing histories than the two-operation chains studied here. Domain-adaptive fine-tuning of the CAS stream on forensic-specific frequency data could further improve performance under extreme compression conditions. Finally, developing saliency-based visualisation tools that highlight the spatial regions driving AFNet’s classification decisions would improve the interpretability of the system and facilitate its adoption by forensic practitioners who need to explain their analysis conclusions to non-expert stakeholders.

BIBLIOGRAPHY

- [1] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.
- [2] X. Chu, Y.-Q. Chen, Q. Liu, and L. Chen, "Detectability of different image manipulation operations," in *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, 2015, pp. 43–48.
- [3] B. Bayar and M. C. Stamm, "Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2691–2706, 2018.
- [4] X. Liao, K. Li, X. Zhu, and K. J. R. Liu, "Robust detection of image operator chain with two-stream convolutional neural network," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 955–968, 2020.
- [5] H. Yuan, W. Luo, and J. Huang, "Blind forensics of median filtering in digital images," *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 4, pp. 1335–1345, 2011.
- [6] M. Boroumand and J. Fridrich, "Deep learning for detecting processing history of images," in *Media Watermarking, Security, and Forensics*, vol. 10579, 2018.
- [7] X. Feng, I. J. Cox, and G. Doerr, "Normalized energy density-based forensic detection of resampled images," *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 536–545, 2012.
- [8] G. Cao, Y. Zhao, R. Ni, and X. Li, "Contrast enhancement-based forensics in digital images," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 3, pp. 515–525, 2014.
- [9] P. Kadha *et al.*, "Remrenet: Residual multi-resolution network for image manipulation chain detection," in *Proceedings of the IEEE International Conference on Image Processing*, 2023.

- [10] L. Yu, Q. Han, and X. Niu, "Manipulation classification for JPEG images using a two-stream network," *IEEE Access*, vol. 8, pp. 27 814–27 823, 2020.
- [11] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Learning rich features for image manipulation detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1053–1061.
- [12] J. Zhong and C.-M. Pun, "An end-to-end dense-inceptionnet for image copy-move forgery detection," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 2134–2144, 2020.
- [13] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.
- [14] D.-T. Dang-Nguyen, C. Pasquini, V. Conotter, and G. Boato, "RAISE: A raw images dataset for digital image forensics," in *Proceedings of the 6th ACM Multimedia Systems Conference*, 2015, pp. 219–224.
- [15] T. Gloe and R. Böhme, "The dresden image database for benchmarking digital image forensics," in *Proceedings of the ACM Symposium on Applied Computing*, 2010, pp. 1584–1590.
- [16] G. Schaefer and M. Stich, "UCID: An uncompressed colour image database," in *Storage and Retrieval Methods and Applications for Multimedia*, vol. 5307, 2004, pp. 472–480.
- [17] A. Rana *et al.*, "MSRD: Multi-scale residual dense network for manipulation chain detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

Forensic Analysis and Detection of Manipulated Images Using Deep Learning Techniques

ORIGINALITY REPORT

8%

SIMILARITY INDEX

7%

INTERNET SOURCES

6%

PUBLICATIONS

%

STUDENT PAPERS

PRIMARY SOURCES

1

arxiv.org

Internet Source

1%

2

export.arxiv.org

Internet Source

1%

3

publications.idiap.ch

Internet Source

<1%

4

www.researchgate.net

Internet Source

<1%

5

cse.hkust.edu.hk

Internet Source

<1%

6

www.frontiersin.org

Internet Source

<1%

7

Geunho Lee, Jieun Lee, Tae-yong Kim, Jongpil Jeong. "An Improved YOLOv9 Based Object Detection with Attention Mechanism for Personal Protective Equipment", Sensors, 2026

Publication

<1%

8

wrap.warwick.ac.uk

Internet Source

<1%

9

techscience.com

Internet Source

<1%

10 Liao, Bowen. "Landslide Change Detection Using Denoising Diffusion Probabilistic Models (DDPM): The Role of Feature Configuration and Topological Integration.", Arizona State University

Publication

11 prod-ms-be.lib.mcmaster.ca <1 %

Internet Source

12 researchspace.ukzn.ac.za <1 %

Internet Source

13 www.mdpi.com <1 %

Internet Source

14 tel.archives-ouvertes.fr <1 %

Internet Source

15 thesis.caltech.edu <1 %

Internet Source

16 www.kresttechnology.com <1 %

Internet Source

17 faui1-files.cs.fau.de <1 %

Internet Source

18 ntnuopen.ntnu.no <1 %

Internet Source

19 Vijayakumar Kadha, V V N J Sri Lakshmi Nandikattu, Sambit Bakshi, Santos Kumar Das. "Forensic analysis of manipulation chains: A deep residual network for detecting JPEG-manipulation-JPEG", Forensic Science International: Digital Investigation, 2023

Publication

research.sabanciuniv.edu

20

Internet Source

<1 %

21

Dongkyu Kim, Han-Ui Jang, Seung-Min Mun, Sunghee Choi, Heung-Kyu Lee. "Median Filtered Image Restoration and Anti-Forensics Using Adversarial Networks", IEEE Signal Processing Letters, 2018

Publication

<1 %

22

Omar Alirr, Razan Alshatti, Shahad Altemeemi, Shahad Alsaad, Alzahraa Alshatti. "Automatic Brain Tumor Segmentation from MRI Scans using U-net Deep Learning", 2023 5th International Conference on Bio-engineering for Smart Technologies (BioSMART), 2023

Publication

<1 %

23

amazingalgorithms.com

Internet Source

<1 %

24

dokumen.pub

Internet Source

<1 %

25

diposit.ub.edu

Internet Source

<1 %

26

Belhassen Bayar, Matthew C. Stamm. "Constrained Convolutional Neural Networks: A New Approach Towards General Purpose Image Manipulation Detection", IEEE Transactions on Information Forensics and Security, 2018

Publication

<1 %

27

Yuanman Li, Jiaxiang You, Jiantao Zhou, Wei Wang, Xin Liao, Xia Li. "Image Operation Chain

<1 %

Detection with Machine Translation
Framework", IEEE Transactions on
Multimedia, 2022

Publication

-
- | | | |
|----|--|------|
| 28 | brage.bibsys.no
Internet Source | <1 % |
|----|--|------|
-
- | | | |
|----|--|------|
| 29 | public-pages-files-2025.frontiersin.org
Internet Source | <1 % |
|----|--|------|
-
- | | | |
|----|--|------|
| 30 | www.diva-portal.org
Internet Source | <1 % |
|----|--|------|
-
- | | | |
|----|---|------|
| 31 | Arvind Dagur, Sohit Agarwal, Dharendra Kumar Shukla, Shabir Ali, Sandhya Sharma.
"Artificial Intelligence and Sustainable Innovation - Volume 1", CRC Press, 2026
Publication | <1 % |
|----|---|------|
-
- | | | |
|----|---|------|
| 32 | Kejun Zhang, Yu Liang, Jianyi Zhang, Zhiqiang Wang, Xinxin Li. "No One Can Escape: A General Approach to Detect Tampered and Generated Image", IEEE Access, 2019
Publication | <1 % |
|----|---|------|
-
- | | | |
|----|--|------|
| 33 | assajournal.com
Internet Source | <1 % |
|----|--|------|
-
- | | | |
|----|--|------|
| 34 | coek.info
Internet Source | <1 % |
|----|--|------|
-
- | | | |
|----|--|------|
| 35 | Anwar, Noreen. "From Temporal Coherence to Cross-Modal Intelligence: A Modular Framework for Video Object Detection.", Ecole Polytechnique, Montreal (Canada)
Publication | <1 % |
|----|--|------|
-
- | | | |
|----|--|------|
| 36 | Iyyakutti Iyappan Ganapathi, Sajid Javed, Syed Sadaf Ali, Arif Mahmood, Ngoc-Son Vu, | <1 % |
|----|--|------|

Naoufel Werghi. "Learning to Localize Image Forgery Using End-to-End Attention Network", Neurocomputing, 2022

Publication

37	Lecture Notes in Electrical Engineering, 2014.	<1 %
----	--	------

Publication

38	Sun, Jian. "Addressing the Problems of Data Variations, Quality, and Scarcity in Training Deep Neural Networks.", University of Denver	<1 %
----	--	------

Publication

39	Wasswa Shafik, Jay Kumar Pandey, Haipeng Liu. "Precision Digital Oncology - Disruptive Science in the Fight Against Cancer", CRC Press, 2026	<1 %
----	--	------

Publication

40	ar5iv.labs.arxiv.org	<1 %
----	---	------

Internet Source

41	content.scirp.org	<1 %
----	---	------

Internet Source

42	deepai.org	<1 %
----	---	------

Internet Source

43	www.medrxiv.org	<1 %
----	---	------

Internet Source

44	www.micc.unifi.it	<1 %
----	---	------

Internet Source

45	www.research.unipd.it	<1 %
----	---	------

Internet Source

46	"Computer Vision – ECCV 2018", Springer Science and Business Media LLC, 2018	<1 %
----	--	------

Publication

47 Md Atiqur Rahman Ahad, Anton Nijholt, Md Abdus Samad Kamal, Björn Schuller, Matthew A. Turk. "Applied Machine Learning on Sensing Technologies", Routledge, 2026
Publication

48 Pushpa Choudhary, Sambit Satpathy, Arvind Dagur, Dharendra Kumar Shukla. "Recent Trends in Intelligent Computing and Communication", CRC Press, 2025
Publication

49 Xiaoying Pan, Nijuan Zhang, Hewei Xie, Shoukun Li, Tong Feng. "MBGNet:Multi-branch boundary generation network with temporal context aggregation for temporal action detection", Applied Intelligence, 2024
Publication

50 Jiahui Cai, Jianguo Hu. "3D RANs: 3D Residual Attention Networks for action recognition", The Visual Computer, 2019
Publication

51 Xiaohan Zhang, Haipeng Wang, Congan Xu, Yafei Lv, Chunlong Fu, Huachao Xiao, You He. "A Lightweight Feature Optimizing Network for Ship Detection in SAR Image", IEEE Access, 2019
Publication
