German International University

Computer Science Faculty

Countering Visual Disinformation: Detection And Localization Of Image Splicing Using Deep Vision Models

Bachelor Thesis

Author:	Fedaa Khalid Abdelghani
Supervisors:	Dr. Nada Sharaf
	Eng. Mohamed Ehab
Submission Date:	18 January, 2024

This is to certify that:

(i) the thesis comprises only my original work toward the Bachelor Degree

(ii) due acknowledgement has been made in the text to all other material used

Fedaa Khalid Abdelghani
18 January, 2024

Acknowledgments

I would like to express my deepest gratitude to Dr. Nada Sharaf for providing me with constant support and positive feedback throughout my journey. I also cannot be more thankful to Eng. Mohamed Ehab, my co-supervisor, for his technical expertise and advice. Lastly, I would like to extend my sincere thanks to Ma Xiaochen from Sichuan University who has generously and selflessly helped me use and learn from his own project to use as foundation in this research.

Abstract

The proliferation of image forgery in the digital age has raised concerns regarding the authenticity and credibility of visual information. This thesis focuses on the detection of forgery in digital images, specifically image splicing. Image splicing is a method of forgery where different sections of multiple images are combined to create a composite image. The thesis proposes two deep vision models: the ELA-DenseNet for image forgery classification and a fine-tuned vision transformer for the localization of manipulated areas. The importance of data quality and quantity is emphasized by introducing an enriched dataset specifically curated and annotated for image splicing detection. The results of the classification model outperform other similar state of the art models and the localization model has shown promising results in tandem with the proposed dataset, thereby advancing the capabilities of forgery detection in this domain and aiding in preserving the trust and integrity of visual information in an era of widespread disinformation.

Contents

Chapter 1

Introduction

The role of images in human communication has evolved significantly, especially with the advent of Web 2.0, which transformed online experiences to be more interactive and participatory [?]. Web 2.0 describes the current state of the internet that focuses on user generated content and participatory culture. This shift allowed for the efficient dissemination of visual information, empowering individuals and spearheading the era of citizen journalism [?]. In this new era, amateur contributors gain the ability to amass exposure and influence comparable to that of more established organizations. While it can mostly be positive, this democratization of image dissemination brought about a concerning issue – the ease of forging information, particularly images. Advanced image processing software such as Photoshop, Corel Paint Shop and GIMP has made it easier than ever to create deceptive visual content, leading to an increase in image forgery and contributing to the current era of disinformation. In this context, disinformation refers to intentionally spreading false information, distinct from misinformation, which is false but is not shared with malicious intent. [?]

The consequences of image forgery extend beyond personal narratives and social media, affecting security and legal contexts where the authenticity of visual evidence is crucial. Manipulated images can be used to construct false narratives, present falsified evidence, or deceive viewers, compromising the credibility of information and distorting public perceptions. Recognizing the gravity of disinformation, particularly emanating from falsified images, the United Nations has identified it as a phenomenon requiring concerted efforts to combat [?], reflecting the concern of 85% of people worldwide, as revealed by a 2023 global survey. [?]

Addressing the challenges of image forgery-induced disinformation involves the application of digital forensics techniques. Digital forensics, a branch of forensic science, involves collecting, analyzing, and preserving digital evidence to investigate and establish the authenticity and integrity of digital artifacts [?]. In image forensics, digital forensics methodologies and tools are used to detect and analyze signs of image manipulation. Leveraging advancements in computer vision, pattern recognition, and machine learning, digital forensics experts can develop robust algorithms to identify forged images, thereby preserving trust and integrity in visual information. [?]

This thesis focuses on a crucial facet of image forgery detection - image splicing detection. Image splicing is a forgery method that involves artfully combining different sections of multiple images to craft a composite image, typically through cutting and pasting [?]. The detection process involves two primary steps. The initial step is a classification task dedicated to determining the authenticity of the image. The second step, termed localization, entails locating the 'forged' portion of the image or predicting its mask—a black-and-white image that designates the tampered area. An illustrative example of an image mask is presented in Figure ??. In tandem, this work's central contribution in the field of image splicing detection arises from the collaborative efforts of two distinct models. One model, the ELA-DenseNet, is dedicated to the classification task, discerning the authenticity of the image through the integration of Error Level Analysis (ELA) and DenseNet architecture. The second model, a fine-tuned vision transformer, specializes in the crucial task of localizing manipulated areas within the image. Together, these two models form a comprehensive approach to address the challenges posed by image splicing, collectively advancing the capabilities of forgery detection in this domain.

In addition to the two models, the work extends the purview of image splicing detection by emphasizing the significance of data quality and quantity. A meticulously expanded dataset tailored for image splicing detection is introduced, encompassing a diverse array of manipulated images. This enriched dataset serves as a valuable resource for training and evaluating the proposed vision transformer and other models, ultimately enhancing generalization and performance in real-world scenarios. The collective efforts of the vision transformer, ELA-DenseNet model, and the enriched dataset collectively propel the capabilities of image forgery detection, particularly in the domain of image splicing.

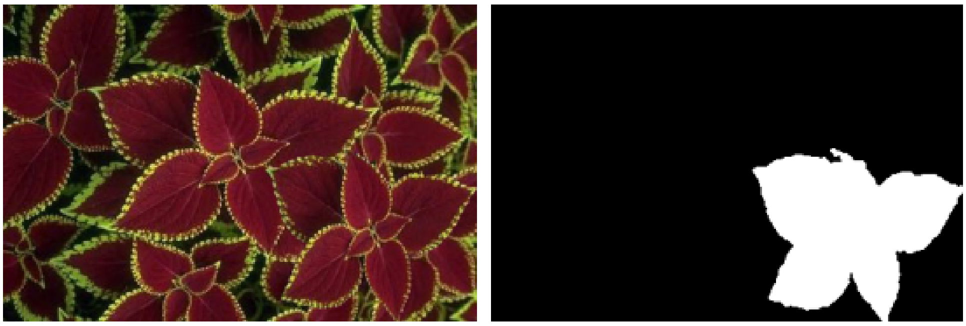

Figure 1.1: An image alongside its image mask showing the tampered region in white [?].

Chapter 2

Background

2.1 Evolution of Image Forgery

Image forgery involves altering the visual contents of an image without leaving traces of the manipulation. This manipulation could be the addition, modification, or removal of any essential features of the image [?]. Image forgery is not a new phenomenon and historically predates the advent of the digital image. In fact, the first forged image in recorded history dates back to the American Civil War in 1860 [?], 34 years after the invention of photography [?], where a photograph of the politician John Calhoun was manipulated and his body was used in another photograph with the head of the president of the United States, Abraham Lincoln. (Figure ??)

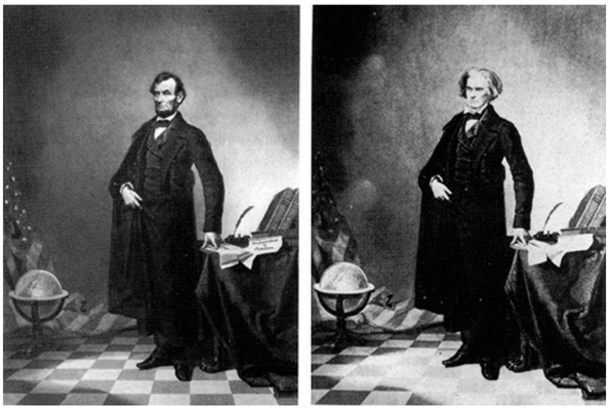

Figure 2.1: The first forged image (left) and the original image (right) [?].

Numerious previous work in the existing literature [?] [?] [?] [?] claim that the first 'forged' photograph was created in 1840 by French photographer and pioneer Hippolyte Bayard, however this is a misconception. Titled 'Self Portrait as a Drowned Man', the image is staged to show him having committed suicide; this was done as an act of frustration towards having lost the title of 'Inventor of Photography' to Louis Daguerre, who patented the process before him (Figure **??**). The error in this claim lies in the fact that this is, in fact, the first 'staged' image and not the first forged image, since the deception occurred before the creation of the image where in contrast forgery typically involves manipulating an existing image.

Figure 2.2: Self Portrait as a Drowned Man, the first staged image. [?]

When it comes to pinpointing the first instance of digital image forgery, the task becomes quite challenging due to the lack of a widely accepted definition for what constitutes a digital forgery during the days of early digital imaging. Early experiments and forgeries in the field of digital art and computer graphics took place in educational and research environments, therefore an accurate answer is difficult to find. Despite there not being an official 'first digital forgery' there are early notable instances, one being the 'O.J. Simpson Time Magazine cover' from 1994 [?], where Simpson's skin tone was allegedly darkened on the magazine's cover to make him appear more menacing. This incident raised ethical concerns about the manipulation of images in journalism.

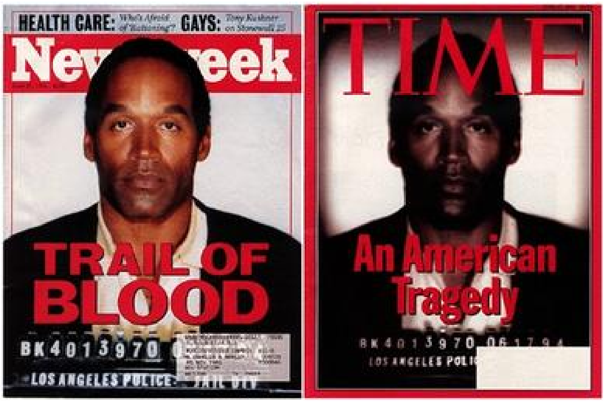

Figure 2.3: O.J. Simpson on the covers of Newsweek and Time Magazine. The picture on the right was altered to make Simpson appear darker. [?]

In the late 20th century, photography greatly evolved into the digital realm, and the advent of graphic imaging software such as Adobe Photoshop or GIMP made digital retouching and editing possible [?]. The skill-intensive and time-consuming nature of using these programs, which were the only options available at the time, made the probability of a proliferation in image forgeries unlikely. Eventually, technical advancements of the 21st century and advent of deep learning created a new wave of image editing software that is powered by advanced algorithms which allow complex transformations to be mostly automated [?]. Nowadays, a number of popular social media apps, like Instagram and TikTok, offer beauty filters which may also be applied in real-time to live video. These filters smooth skin tones and generate more visually pleasing face proportions (for example, by enlarging a subject's eyes). Applications for mobile image editing, such as Facetune, also offer these features. Some, like FaceApp, automate intricate, content-aware alterations, such altering the subject's age or gender or their facial expression, using deep learning algorithms [?]. The ease of use and availability of these technologies, often for free, led to a distrust in the authenticity of visual information, especially on social media.[?] The proliferation of image forgeries was no longer a matter of probability, but became a reality.

The year 2017 saw the emergence and spread of 'Deepfakes', a form of synthetic media in which artificial intelligence (AI) and deep learning techniques are used to create or manipulate audio and visual content to make it appear as though someone is saying or doing something they did not [?]. The term 'deepfake' originated from a Reddit user called 'deepfake' who was known for sharing the deepfakes he created; many videos involved celebrities' faces swapped onto the bodies of actresses in pornographic videos [?]. The term is a portmanteau of 'deep learning' and 'fake'. One major problem is how deepfakes contribute to fake news. The potential to produce extremely convincing and misleading videos increases the likelihood of misinformation and disinformation spreading [?]. Deepfakes create the impression that prominent individuals are saying or doing things they never did, which can be exploited to sway public opinion, harm people's reputations, or affect political outcomes. By 2020, audio deepfakes had emerged [?], and AI software capable of both creating and detecting deepfakes, including cloning human voices after just five seconds of listening time, became available. Impressions, a mobile deepfake app, launched in March 2020, pioneering the creation of celebrity deepfake videos directly from mobile phones [?].

As briefly explored above, image manipulation has been used to deceive or persuade viewers or, more lightheartedly, improve storytelling and self-expression. It then becomes clear that image manipulation is not inherently malicious and it is, in fact, the intent, the nature of the manipulation, and its consequences that determine whether the manipulation can be considered a malicious forgery. This, in turn, allows us to create a consistent definition of what forgeries are and how to classify them, aiding in the task of detecting them and mitigating their spread.

According to a survey done by Meena and Tyagi [?], the act of image forgery itself can be classified into two types: content-preserving or content-altering forgery. Content-preserving forgery includes operations such as geometric transformation, changing the format, enhancement, etc. Content-altering forgery involves altering the content of digital images to misinterpret and mislead the information conveyed by the image, ie. its semantic meaning. As such, one form of tampering can be seen as more malicious than the other. Retouching and resampling are examples of legitimate content-preserving tampering that tries to improve visual appeal or resize photos for specific purposes like magazine covers or social media. Most of the time, these modifications are generally accepted and harmless. On the other hand, content-altering tampering is seen as malicious and illegal tampering as it attacks an image's semantic integrity.

Traditionally, the two primary types of illegal tampering, in which parts are joined or altered to purposefully trick and mislead viewers, are copy-move forgery and image splicing forgery [?]. Copy-move forgery involves duplicating and placing one part of an image onto another to create a false appearance of objects being in multiple locations. Image splicing, as shown in Figure ??, is a manipulation technique where different parts of multiple images are combined to create a deceptive composite image. Deepfakes can be considered the new, third type of illegal tampering as they do not fall under the categories of either copy-move or image splicing.

2.2 Digital Image Forensics

Image forgery detection refers to the process of identifying manipulations or alterations in digital images that are intended to deceive or mislead. The goal is to distinguish between authentic images and those that have been tampered with, ensuring the integrity and reliability of visual content [?]. Image forgery detection techniques are broadly classified into two categories: active and passive. Active forgery detection requires prior embedded information about the query image, which can be added during image acquisition or later using hardware or software. The embedded data is then extracted from the query image to assess the likelihood of forgery. This data is mostly embedded in the form of a watermark or a digital signature.[?] However, active forgery detection has limitations when applied to photographs from the Internet or social media, as these images lack embedded information. In response, passive forgery detection methods have been developed, operating without prior knowledge of the query image.[?]

Digital forgeries leave almost no visual clues with regard to tampering, but the image structure is disturbed, which introduces new artifacts resulting in various forms of inconsistencies [?]. Passive methods, also known as Digital Image Forensics (DIF), rely on the application of forensic techniques to analyze these inconsistencies in the digital image to detect the forgery[?]. Image forgery detection is a two-class (binary) classification problem. The objective of a DIF model is to classify a given image as either authentic or forged. As shown in Figure ??, the general framework of the existing DIF approaches includes extracting representative and relevant features from an image first, then a suitable classifier is trained and modeled by using the features, and finally classification is performed by using the trained model.

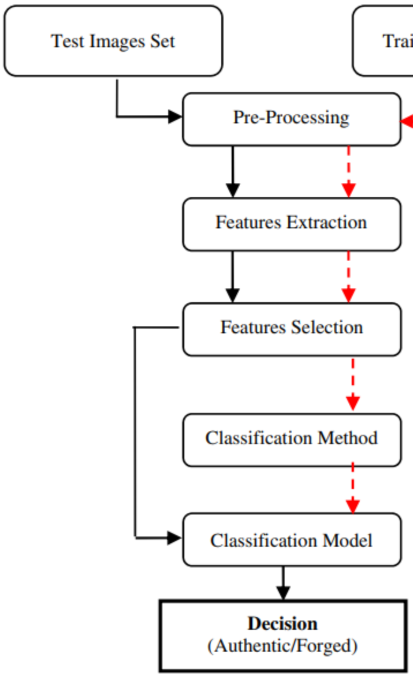

Figure 2.4: General Framework for forgery detection.

Figure ?? shows how the passive DIF is categorized into forgery-dependent approaches and forgery-independent approaches.Forgery-independent techniques concentrate on finding irregularities or inconsistencies within the image itself, while forgery-dependent techniques rely on particular forgery features and assumptions about the alteration process [?]. Forgery-independent techniques by definition have a wider use but may have some limits in terms of false positives or the identification of complex forgeries, meanwhile forgery-dependent strategies can be effective when the forgery process is known.

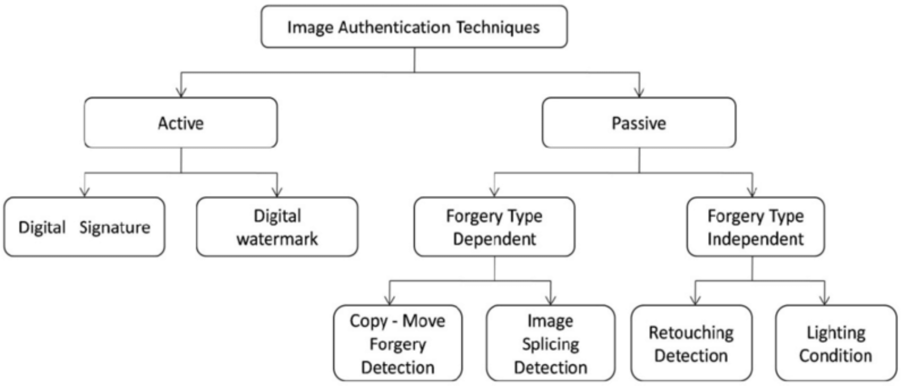

Figure 2.5: Classification of image authentication techniques. [?]

As mentioned in the previous section, the three main types of illegal tampering are copy-move, image splicing and deepfakes. Consequently, the majority of effort and research done in the field of DIF is focused on these forgery techniques [?]. However, as observed in many review and survey papers [?], there are relatively few approaches dedicated to detecting image splicing compared to copy-move forgery detection (CMFD) and deepfakes; this represents a gap in the research, and therefore, image splicing detection techniques are the focus of this thesis and will be highlighted in the related works section.

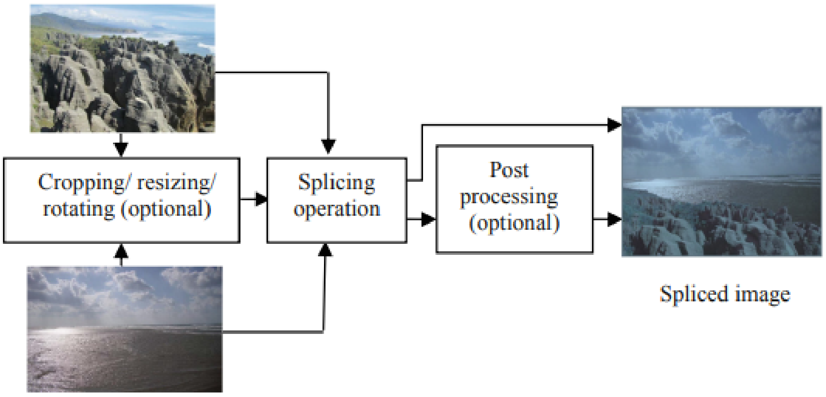
Cropping/ resizing/ rotating (optional)
Splicing operation
Post processing (optional)
Spliced image

Figure 2.6: Steps to creating a spliced image. [?]

2.3 JPEG Compression & Image Forensics

JPEG (Joint Photographic Experts Group) stands as a widely embraced image compression standard, celebrated for its efficiency in storing and transmitting digital images. Its popularity stems from the ability to achieve notable compression ratios while preserving acceptable image quality. [?] This algorithm, widely utilized in various applications such as digital photography, web content, and social media, achieves compression by removing redundant information and approximating image data. The adaptability of JPEG, allowing for an adjustable tradeoff between storage size and image quality, further cements its dominance in the digital imaging landscape, supported by widespread compatibility [?].

Being the default and most widely supported format for image storage and sharing on the internet, JPEG plays a pivotal role in the digital realm. The majority of digital cameras, smartphones, and image-sharing platforms employ JPEG compression, making it integral to the fabric of online visual content.

However, this widespread usage has inadvertently led to concerns related to image forgery, especially in the context of image splicing. JPEG compression, while achieving efficient file sizes, introduces artifacts during the compression process. Lossy compression sacrifices specific image details and may result in the smoothing of high-frequency components. The Discrete Cosine Transform (DCT) and blocking artifacts at block boundaries, coupled with quantization errors, contribute to distinctive patterns within the image. Forgers adeptly utilize these compression artifacts to conceal seams between spliced regions during image manipulation [?]. This strategic use makes it challenging for traditional forgery detection methods to discern the manipulation. Consequently, advanced splicing techniques are continually evolving to effectively conceal these artifacts, fueling ongoing research into sophisticated forgery detection methods.

The ubiquity of JPEG images on social media platforms renders them susceptible to the dissemination of manipulated or forged images, accentuating the potential for the spread of disinformation. Disinformation campaigns leverage manipulated images to shape public perception, spread false narratives, and create misleading visual evidence. JPEG splicing, enabling the creation of realistic yet fabricated images, stands as a potent tool for deception and misinformation in this digital age [?]. As a result, the detection of forgeries in JPEG images especially is a thriving branch of digital image forensics and therefore, the contributions of this work are mainly targeted at addressing the splicing of JPEG images.

2.4 Deep Learning in Digital Image Forensics

Machine learning, a subset of AI, empowers systems to learn and improve from experience without explicit programming. Deep learning is a branch of machine learning that involves the use of neural networks with multiple layers (known as deep neural networks) to analyze and learn patterns from vast amounts of data [?]. The journey of deep learning commenced with the development of artificial neural networks, inspired by the structure and functioning of the human brain. These networks are capable of automatically extracting hierarchical features and representations, making them highly effective in tasks such as image recognition, classification, and pattern detection [?].

The use of deep learning in image forensics represents a more recent development, leveraging the capabilities of these advanced algorithms to detect and analyze digital image manipulations. The integration of deep learning techniques in image forensics gained traction as computational power increased, allowing for the training of complex neural networks on large datasets [?].

2.4.1 Convolutional Neural Networks

The concept of convolutional neural networks (CNNs) is inspired by the human visual system. It was introduced in a landmark paper titled ‘Gradient-Based Learning Applied to Document Recognition’ by Yann LeCun et. al in 1998 [?]. The paper proposed the LeNet-5 architecture, specifically designed to recognize handwritten digits in checks and postal codes, demonstrating the effectiveness of CNNs in learning hierarchical features from input data. This work marked a groundbreaking milestone in the history of deep learning and became foundational for subsequent applications in computer vision, including image forensics.

CNNs revolutionized computer vision by addressing the spatial hierarchies within images. The architecture of CNNs mimics the human visual system, employing convolutional layers to recognize local patterns, pooling layers to downsample, and fully connected layers for classification [?]. The success of CNNs in tasks like image recognition, object detection, and segmentation solidified their status as the go-to architecture in the realm of computer vision [?]. These networks excel at feature extraction, capturing hierarchical representations of input data. The convolutional layers apply filters to detect low-level features like edges and textures, gradually combining them to identify more complex patterns [?]. CNNs' efficacy in image classification is exemplified by their deployment in renowned architectures like AlexNet [?], VGGNet [?], and ResNet [?], setting benchmarks in various computer vision competitions.

While CNNs thrived, their reliance on fixed-size grids and limited global context became apparent challenges [?]. The breakthrough in natural language processing with transformers, as showcased by models like BERT and GPT, inspired researchers to explore their potential in computer vision [?].

2.4.2 Vision Transformers

Vision Transformers (ViTs) emerged as a groundbreaking architecture for computer vision, challenging the conventional dominance of Convolutional Neural Networks (CNNs). The ViT model, introduced in the paper 'An Image is Worth 16x16 Words: Transformers for Image Recognition' by Alexey Dosovitskiy et al., marked a significant departure from grid-based processing, treating images as sequences of fixed-size patches [?].

At the heart of ViTs is the transformer architecture, initially designed for natural language processing tasks. ViTs, however, adapt this architecture to handle visual data by dividing an image into non-overlapping patches, which are then linearly embedded into high-dimensional vectors. These patch embeddings serve as the input tokens for the transformer model [?].

The transformer consists of two main components: the encoder and the decoder. In ViTs, only the encoder is used [?]. The encoder comprises multiple layers, each containing self-attention mechanisms and feedforward neural networks. The self-attention mechanism enables the model to capture relationships between different patches, allowing for the extraction of global contextual information. The model learns to attend to relevant patches while suppressing irrelevant ones, effectively achieving a holistic understanding of the image [?].

Detailed Structure of Vision Transformers:

1. Input Embedding: The input image is divided into non-overlapping patches, and each patch is linearly embedded into a flat vector. These patch embeddings serve as the input to the transformer.

2. Positional Embedding: To maintain spatial information, positional embeddings are added to the patch embeddings. This helps the model understand the relative positions of different patches in the image.

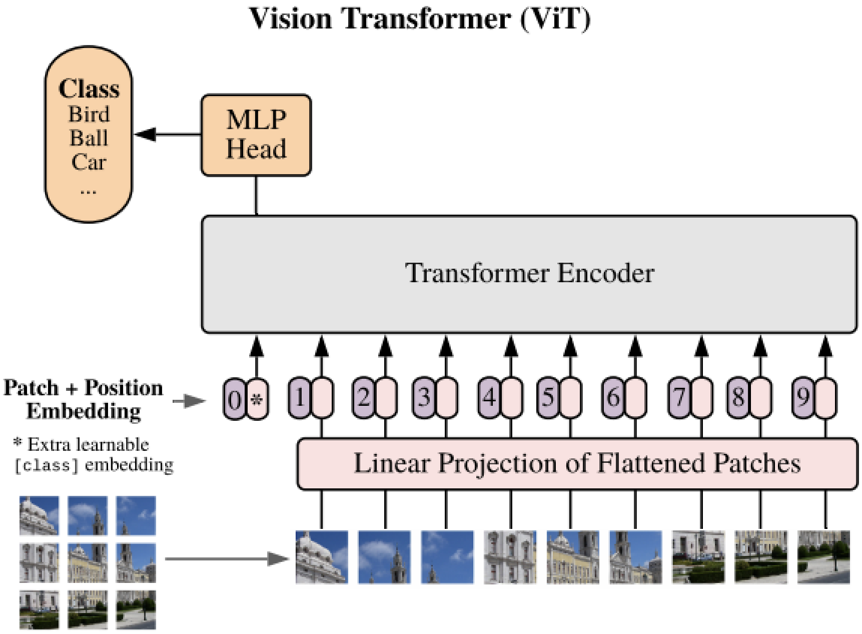
Vision Transformer (ViT)
Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0
*
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches

Figure 2.7: Architecture of Vision Transformers [?].

3. Transformer Encoder: The core of ViTs is the Transformer Encoder, consisting of multiple identical layers. Each layer has two main components: self-attention mechanism and a feedforward neural network.

4. Self-Attention Mechanism: This mechanism allows each patch to attend to all other patches, capturing long-range dependencies in the image. Self-attention helps the model understand the relationships between different parts of the image.

5. Feedforward Neural Network: After self-attention, the patch embeddings pass through a multi-layer perceptron (MLP) or feedforward neural network. This MLP introduces non-linearity and allows the model to learn complex representations.

6. Layer Normalization and Residual Connections: Each sub-layer (self-attention and feedforward) is followed by layer normalization, and the output is then passed through a residual connection. These components help stabilize training.

7. Encoder Stack: Multiple transformer encoder layers are stacked on top of each other. The sequential application of these layers allows the model to capture hierarchical features.

8. Global Average Pooling: After the encoder stack, the output is typically processed through a global average pooling layer, reducing the spatial dimensions to a single vector for classification tasks.

9. MLP Head: At the end of the ViT architecture, a Multi-Layer Perceptron (MLP) head is added for task-specific processing. This MLP head takes the output of the global average pooling and transforms it to the desired output format (e.g., class scores for classification tasks) [?].

ViTs offer several advantages that distinguish them from traditional CNNs. Firstly, ViTs excel in capturing long-range dependencies in images, a feat beyond the limited receptive fields of CNNs [?]. By considering relationships between all patches, ViTs can comprehend global contextual information, proving invaluable for tasks requiring a holistic understanding of the input. Additionally, ViTs exhibit remarkable versatility, adeptly handling various computer vision tasks such as image classification, object detection, and segmentation [?] [?]. Their adaptable architecture contributes to their effectiveness across a spectrum of visual tasks. Unlike traditional CNNs, ViTs mitigate grid dependencies by treating images as sequences rather than relying on grid structures and hierarchical representations [?]. This departure enables ViTs to excel in tasks where grid-based approaches might falter. Moreover, ViTs leverage transfer learning through pre-training on large-scale datasets, akin to their counterparts in natural language processing. This strategy empowers ViTs to learn rich, generalized representations, facilitating fine-tuning for specific tasks using smaller datasets [?]. Collectively, these advantages position Vision Transformers as a significant breakthrough in the realm of deep learning for computer vision.

2.5 Localization and Semantic Segmentation

In the realm of computer vision, localization and segmentation are pivotal tasks that enhance our comprehension of visual content [?]. Localization, also known as detection, entails pinpointing the exact location of objects within an image, commonly achieved through the prediction of bounding boxes around these objects. Conversely, segmentation takes a more intricate approach, providing pixel-level accuracy in delineating object boundaries and generating detailed masks that distinguish various entities within the image [?]. The chronological evolution of these tasks is noteworthy. Object localization, centered on predicting bounding boxes, emerged as an initial focus in computer vision research. This foundational work set the stage for subsequent advancements, including the development of more sophisticated object detection and segmentation methods.

Segmentation involves dividing an image into meaningful segments or regions. More precisely, segmentation can be divided into two types: semantic segmentation and instance segmentation. Semantic segmentation treats multiple objects within a single category as one entity. Instance segmentation, on the other hand, identifies individual objects within these categories [?]. Within the realm of image forgery detection, semantic segmentation dominates the field as the segmentation task requires no further instance evaluation, in the sense that we only care about the tampered category and less about classifying the contents within it. The difference between the two is highlighted in Figure ?? Traditional semantic segmentation techniques included thresholding, edge-based

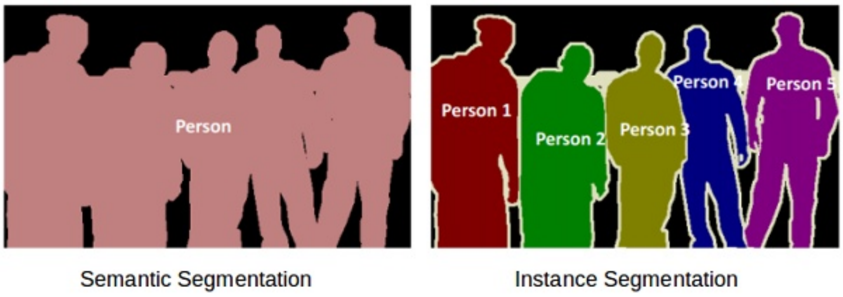

Figure 2.8: Semantic and Instance Segmentation. [?]

methods, and region-growing algorithms [?]. Thresholding assigns pixel values based on predefined intensity thresholds, while edge-based methods identify boundaries through gradient analysis [?]. Region-growing algorithms group adjacent pixels with similar characteristics. Deep learning brought about a transformative shift in segmentation with Fully Convolutional Networks (FCNs) leading the way. FCNs utilize convolutional layers to capture spatial information and upsampling layers to generate dense pixel-wise predictions [?]. U-Net, a popular architecture, combines a contracting path for context extraction and an expansive path for precise localization [?].

Recent state-of-the-art segmentation models often leverage the power of attention mechanisms. ViTs have been adapted for computer vision tasks, proving particularly effective for segmentation as they process images in the form of patch sequences, capturing global contextual information and enhancing segmentation accuracy [?] [?].

In essence, while object localization initially addressed the question of ‘where’ in computer vision, segmentation has expanded this understanding to the level of ‘what’ and ‘how much’ by delineating object boundaries at a finer granularity. The progression from bounding boxes to pixel-wise predictions showcases the iterative development of localization techniques that have influenced and contributed to the evolution of segmentation methods. In image forgery detection, these tasks become pivotal for unveiling the authenticity of images. Localization helps identify tampered regions within an image, while segmentation aids in precisely delineating the forged portions. By combining both techniques, one can achieve a comprehensive understanding of the manipulated areas, contributing to the development of robust forgery detection algorithms.

In the existing literature, there is an almost unanimous tendency to employ the term ‘localization’ rather than ‘segmentation’ when describing the methods applied in image forensics [?] [?] [?] [?] [?] [?] [?]. This linguistic choice often arises from a desire to convey a sense of pinpointing specific regions of interest within an image, while also encompassing the broader objective of identifying manipulated areas. While the true technical process involves segmentation, where regions are precisely delineated and differentiated, the term ‘localization’ is employed to emphasize the goal of isolating and detecting forged regions without necessarily implying a detailed pixel-wise separation. This semantic shift in language reflects a pragmatic approach, simplifying the terminology for a wider audience and facilitating a more intuitive understanding of the primary objective - detecting and localizing image forgeries [?]. Following this approach, for alignment with the consensus, the term ‘localization’ is used in this thesis in the same manner.

Chapter 3

Related Work

Image splicing detection has evolved through various methodologies, including traditional techniques, deep learning approaches, and splicing localization methods [?]. This chapter explores the wide range of relevant works and charts the developments made in the field of image splicing research. Firstly, the discussion unfolds around early traditional techniques comprising handcrafted feature-based methods. Then we explore contemporary methods levering deep learning architectures. Finally, we highlight the efforts dedicated to the task of localizing the spliced regions.

3.1 Handcrafted feature-based methods

In this section, we explore traditional image splicing detection methods, also referred to as handcrafted feature-based methods. These methods involve the manual creation and extraction of specific features from images, such as texture patterns, color variations, and statistical properties, and relying on inconsistencies within them to identify manipulations [?]. We explore the most commonly [?] used methods which are:

1. Noise-based Methods
2. Textural feature-based Methods
3. Markov feature-based Methods

3.1.1 Noise-based Methods

Noise-based methods identify manipulations by analyzing inconsistencies in noise patterns. Digital images acquire noise during capture, stemming from light (physical noise) and sensor imperfections (hardware noise). Images captured on different cameras will have inherently different noise levels, resulting in spliced photos displaying uneven noise patterns [?]. Various methods have been devised to identify image splicing forgeries by utilizing noise variations. Meena and Tyagi [?] introduced a noise level estimation method that utilizes irregular-shaped superpixel blocks, employing the Simple Linear Iterative Clustering (SLIC) technique. The approach relies on principal component analysis (PCA) for image estimation and k-means clustering to differentiate between authentic and spliced blocks based on noise levels. Another approach by Itier et. al [?] focuses on color noise correlation analysis across RGB channels, exploiting intrinsic features of image acquisition to detect spliced areas by identifying blocks spanning both the background and spliced regions. Additionally, a PCA noise estimation method, proposed by Zhan et. al [?], capitalizes on variations in noise levels across images from different sources, identifying inconsistencies in local noise levels. The study [?] suggested a noise pattern analysis method, segmenting images into non-overlapping blocks and estimating noise variances through wavelet transformations on the luminance component of the image.

Error Level Analysis

A tool called error level analysis (ELA), which can evaluate images with different levels of compression, was proposed in a study [?]. In ELA, an image is first compressed at a known error rate, and then the difference between the original and compressed versions is analyzed. Regions that exhibit higher error levels than the surrounding areas are considered potential areas of interest for further investigation. ELA is based on the premise that authentic images typically have consistent error levels across the entire image, while manipulated regions may show variations in error levels. ELA specifically ties into the detection of forgeries in JPEG images. JPEG compression introduces quantization errors, and these errors are more pronounced in regions of the image that have been modified or manipulated. ELA takes advantage of these errors to highlight potential areas of interest for further investigation [?]. [?] examines and demonstrates the great reliability of this technique for detecting image splicing. A method for forgery detection in lossy compressed digital images employing ELA and filtering its noise components using automatic wavelet soft thresholds was proposed in 2019 [?]. More recently, a study combined ELA with the Local Binary Pattern (LBP) for splicing detection, achieving an accuracy score of 91.46% on the Columbia dataset, a benchmark dataset consisting of 363 images [?].

3.1.2 Textural Feature-Based Methods

Textural Feature-Based Methods focus on texture analysis to detect splicing. Texture analysis attempts to quantify texture qualities described by terms such as rough, smooth, silky, or bumpy as a function of the spatial variation in pixel intensities [?]. In image splicing forgery, new micro-patterns replace the original ones in the spliced area, making edges sharp along the spliced region boundaries. Consequently, the local frequency distributions in the spliced image notably differ from those in an authentic image. Additionally, each source image contributing to a composite image may have a distinct spatial arrangement of colors or intensities [?]. Alahmadi et al. [?] proposed a method using LBP and DCT, emphasizing its efficiency in detecting both copy-move and splicing forgeries. Sharma and Ghanekar [?] proposed an algorithm based on Local Directional Pattern (LDP), unique for its combined ability to classify spliced images and localize the specific tampered regions. Srivastava and Yadav [?] presented an approach utilizing Enhanced Local Ternary Pattern (ELTP) for feature extraction, surpassing the performance of Local Binary Pattern (LBP). Following this, Kanwal et al. [?] introduced a novel block-based approach leveraging the Otsu-based Enhanced Local Ternary Pattern (OELTP) for feature extraction, enhancing the forgery detection capabilities of the existing Enhanced Local Ternary Pattern (ELTP).

3.1.3 Markov Feature-Based Methods

Markov Feature-Based methods analyze pixel dependencies, considering not only individual pixel characteristics but also their interactions with neighboring pixels, enabling the detection of subtle patterns and inconsistencies that may not be apparent when analyzing isolated features [?]. Because of their ability to capture these nuances, they have been widely studied [?] and achieved overwhelming success. Rhhman et al. [?] conducted a comparative analysis of image splicing algorithms, introducing an enhanced technique that combines Discrete Cosine Transform Domain (DTC) with Markov features in the spatial domain. The authors utilize Principal Component Analysis (PCA) to select significant features and apply Support Vector Machine (SVM) for classification on a publicly available dataset using ten-fold cross-validation. Yildirim [?] presents a method for image splicing detection with Discrete Wavelet Transform (DWT) domain extended Markov features, utilizing a combination of DWT and Markov-based methods for feature extraction, followed by Support Vector Machine classification. Furthermore, Pham et al. [?] introduced an efficient image splicing detection algorithm based on Markov features. Their method involves applying Discrete Wavelet Transform to the image and extracting four-direction Markov features for subsequent classification using Support Vector Machines. Yildirim & Ulutaş [?] proposed a Markov-based method in the Discrete Cosine Transform domain to detect image splicing forgery. They extract Markov features from high-frequency coefficients obtained through DCT and employ Support Vector Machines for image classification. Notably, they achieved the highest accuracy rate of 99.98%, underscoring the efficacy of Markov-based forgery detection methods across diverse domains and methodologies. However, Markov models suffer from the complexity of calculations and time consumption.

Through this discussion, the importance of traditional and handcrafted feature-based methods and their contributions and limitations become clear, setting the stage for an exploration of more contemporary and automated approaches in the following section.

3.2 Deep Learning Methods

Traditionally, image characteristics were extracted by employing various filters or descriptors. However, as the diversity of features within an image grows, the task of designing filters to extract each type of feature becomes increasingly complex [?]. Moreover, relying on expert opinions to select suitable filters for highlighting specific aspects adds an additional layer of challenge. The emergence of deep learning has effectively addressed these limitations, offering the capability for automatic learning of intricate feature representations [?]. In this section, we focus on the current state-of-the-art technology for forgery detection: CNNs.

In recent years, the landscape of image splicing forgery detection has witnessed a surge in methods leveraging deep learning techniques. Prominent among these frameworks are Deep Boltzmann Machines [?], Deep Autoencoders [?], and Convolutional Neural Networks (CNNs) [?, ?, ?]. Demonstrating impressive performance across various artificial intelligence tasks, these deep learning architectures have served as foundational elements for more advanced models.

Prior to the widespread adoption of CNNs, early endeavors in splicing detection and localization predominantly employed autoencoders. Functioning as neural networks that utilize fewer bits to reconstruct an input image, autoencoders have played a crucial role in identifying and localizing tampered regions on a patch-wise basis. These networks utilize wavelet features as input and can incorporate additional local noise features, such as the Spatial Rich Model (SRM), widely used in steganalysis—the study and detection of hidden messages in digital media [?]. SRM employs manually crafted filters to extract nearby pixel-level noise, facilitating the differentiation between manipulated and original regions. Notably, SRM filters have found application in creating specialized inputs for CNNs. In [?], authors used SRM features as input for their autoencoder model. Futhermore, in [?], Rao and Ni proposed to use 30 SRM filters as initialization for the first layer in a CNN to detect splicing and copy-move forgeries. The results from the pre-trained CNN were used in an SVM classifier for determining whether the images were forged.

3.2.1 Convolutional Neural Networks

CNN, particularly effective in discovering spatial correlations, is able to combine image segmentation, feature extraction, and classification, to provide an automatic and general approach for authenticating or detecting forged images [?], [?], [?], [?]. This intelligent application of deep learning helps model and identify structural changes indicative of forgery, offering a robust solution to the challenges in forgery detection.

Rao et al. [?] introduced a deep learning-based method for image splicing detection, employing a convolutional neural network (CNN) to automatically learn hierarchical features from input images. The approach incorporates high-pass filters to extract noise residuals and utilizes an SVM classifier for further classification. In a similar vein, Liu et al. [?] devised a deep learning model for image splicing forgery detection, incorporating a conditional random field (CRF) to adaptively integrate results from CNNs. CRFs are probabilistic graphical models used to model relationships between pixels in an image, particularly useful for refining and post-processing. The authors reported superior performance compared to existing methods. Pomari et al. [?] proposed a method combining illumination inconsistencies and deep learning to reveal image splicing. Subsequently, Chen et al. [?] enhanced the methodology introduced by Liu et al. [?], introducing a novel approach using three fully convolutional networks (FCNs) with distinct upsampling layers. Furthermore, the authors initialized these FCNs with pretrained weights from the VGG-16 network, contributing to improved performance in image splicing detection. FCNs differ from traditional CNNs by replacing fully connected layers with convolutional layers, enabling pixel-wise predictions and making them well-suited for tasks like image segmentation.

Ahmed et al. [?] introduced a new backbone architecture for deep learning called ResNet-conv. ResNet-conv is obtained by replacing the feature pyramid network (FPN) in Residual Network (ResNet)-FPN with a set of convolutional layers. This new backbone is used to generate the initial feature map, which is then used to train the Mask-Region-based Convolutional Neural Network (RCNN) to generate masks for spliced regions in forged images. Recently, Guo et al. [?] developed a novel hierarchical fine-grained formulation for Image Forgery Detection and Localization (IFDL), utilizing multiple labels at different levels to capture hierarchical dependencies and improve performance across seven benchmarks. Tyagi and Yadav [?] presented a compact Convolutional Neural Network (CNN) tailored for image forgery detection, showcasing efficiency in feature extraction and authenticity classification on both the CASIA V2 and COVERAGE datasets.

3.3 Localization of Image Splicing

As previously stated, the second step to an image forgery detection method is the localization step, whereby the region of tampering is highlighted, traditionally in the form of a black and white mask. Since forgery localization goes hand in hand with forgery detection, it has also gone through the familiar evolution from traditional handcrafted features to deep learning methods, with ViTs being a rapidly emerging technology in forgery localization. The contributions that led to this development will be explored in this section.

Several traditional approaches [?], [?], [?] have utilized JPEG blocking artifacts to detect tampered regions. As previously mentioned, JPEG, as previously mentioned, is a lossy compression that divides an image into blocks and applies DCT to each block, followed by quantization. The process introduces specific blocking artifacts, which can be exploited for tampering detection and localization, utilizing quantitative measures, such as the Mean Squared Error (MSE) or Structural Similarity Index (SSI), to assess the blockiness introduced by splicing. Sharma and Ghanekar [?] tackled the task of localization through splitting the chrominance component of the query image into overlapping blocks and calculating the Local Directional Pattern (LDP) of each block. The standard deviation of each block is then used as a clue to visualize the spliced region.

The introduction of deep learning approaches into the forgery detection field meant that traditional approaches could be enhanced using new technologies; [?] developed an improved Mask R-CNN which attaches a Sobel filter to the mask branch of the Mask R-CNN. The authors use Mask-RCNN to localize the forgery for its unique ability to identify and segment objects at the pixel level. The Sobel filter acts as an auxiliary task to encourage predicted masks to have similar image gradients to the ground-truth mask. Similarly, Rao et al. [?] designed and implemented a CNN with the first layer of the network initialized with 30 SRM filters to locate splicing forgeries that were meant to improve on their previous work [?]. Based on the pre-trained CNN model, an image splicing localization scheme was developed by incorporating the fully connected Conditional Random Field (CRF), which has shown superior performance in semantic segmentation under the framework of deep learning [?]. In [?], a Siamese CNN with a self-consistency approach to determine if contents had been produced by a single device was proposed. The proposed model could predict the probability that two patches had similar Exchangeable Image File (EXIF) attributes and output a heatmap, highlighting image regions that had undergone possible forgery. Bunk et al. [?] proposed a hybrid CNN-Long Short Term Memory (LSTM) model that captures the boundary discrepancy, or spatial structure, between manipulated and non-manipulated regions. The model was trained end-to-end using ground-truth mask information.

3.3.1 Vision Transformers

One of the earliest works in this realm, to the best of our knowledge, was a 2021 paper that proposed a technique that utilizes a ViT to localize manipulated areas within satellite images [?]. The method relies heavily on a basic ViT architecture and requires no labeled data, making it cost effective and a scalable solution. To reduce the memory used by the self-attention module in the transformer, they used the Linformer [?] to reduce the space complexity from the original ViT from quadratic complexity to linear. The authors also created their own dataset of satellite images to mitigate the issue of training data quality.

A more general-purpose method called TransForensics [?] was developed to localize manipulations in regular images. The architecture uses dense self-attention encoders and dense correction modules to model global context and interactions between local patches at different scales. TransForensics is able to capture discriminative representations and generate high-quality mask predictions for various types of tampering, regardless of the order of patch sequences. TransForensics proved to outperform state-of-the-art models, mostly CNNs, on datasets such as CASIA, COVER, and IMD2020. The results of the proposed method were compared with various state-of-the-art models on these datasets and reported an outperforming F1 score of 0.884. ObjectFormer was a proposed method for detecting and localizing image manipulations [?]. The authors address the challenge of detecting subtle manipulation traces that are no longer visible in the RGB domain by using CNNs to extract high-frequency features from images and combining them with RGB features as multimodal patch embeddings. The embeddings are then passed through a ViT for further encoding. The model was evaluated on the CASIA, Columbia, Coverage, NIST16, and IMD2020 datasets, outperforming state-of-the-art approaches with an F1-score of 0.973.

The IML-ViT model introduced in 2023 presents a novel approach addressing the challenging task of Image Manipulation Localization (IML) [?]. By leveraging the self-attention mechanism in ViTs, the authors propose IML-ViT as a high-resolution, multi-scale model capable of capturing artifacts and detecting image manipulations. The paper highlights the limitations of existing CNN-based methods in handling non-semantic modeling and long-range dependencies, making a compelling case for the suitability of ViTs in IML. The authors pointed out the shortcomings of the existing ViT approaches, namely ObjectFormer [?] and TransForensics [?]; both approaches miss important first-hand low-level information since they use multiple CNN layers to first extract feature maps and then use transformers for additional encoding. Therefore, they aimed to simplify the architecture and introduced a morphology-based edge loss strategy to guide the model's focus on the boundaries of manipulated regions, further enhancing its localization performance. Through extensive experiments on benchmark datasets, such as CASIA V1 and Columbia, IML-ViT demonstrates superior performance compared to state-of-the-art methods in terms of accuracy and localization metrics, with the best F1-score being 0.658.

This thesis builds upon existing related work in the field, particularly [?] and [?], aiming to address key gaps in current approaches to forgery detection. Specifically, it focuses on enhancing the generalization of deep models, revitalizing forensic techniques, and elevating data quality and availability. Firstly, recognizing the limitations of hand-crafted feature-based methods, the thesis explores more automated and generalized techniques. Leveraging the power of the new and emerging ViTs, the research tailors these advancements to the unique challenges posed by online image manipulation. Secondly, the thesis proposes integrating ELA, an underexplored forensic technique, with DenseNet architecture to create the ELA-DenseNet model. This combination aims to improve the classification accuracy of forged images. Lastly, the research emphasizes the paramount importance of data quality and quantity. It introduces a an expanded dataset specifically curated and annotated for image splicing detection, strategically aligned with ViT's goal for enhanced performance in online forgery detection.

Chapter 4

Methodology

4.1 Proposed Models

In this section, we present three key contributions aimed at enhancing image splicing detection. Firstly, we introduce the ELA-DenseNet model, a novel fusion of the ELA forensic technique and DenseNet-121 architecture, providing a robust forgery detection system. Secondly, a fine-tuned vision transformer is incorporated to detect and localize manipulated regions in spliced images, harnessing the latest advancements in vision transformer technology. Lastly, a curated dataset of spliced images and corresponding groundtruths is introduced, addressing the critical need for high-quality training and evaluation data. Together, these contributions form a comprehensive and innovative approach to advance image splicing detection, bridging traditional methods with state-of-the-art technologies.

The decision to develop separate models for forgery classification and localization, rather than a unified model for both tasks, is worth noting. It was a deliberate choice driven by a nuanced understanding of the distinct technical demands inherent to each task. Classification entails a comprehensive assessment of the entire image, requiring a model capable of making a binary decision — forged or authentic. This demands the capture of global features and contextual information to discern the overall integrity, aligning well with architectures that are optimized for image classification like CNNs, as exemplified in the ELA-DenseNet model. Conversely, localization zeroes in on specific regions within an image that have undergone manipulation, necessitating a fine-grained analysis. For this task, leveraging object detection and segmentation technologies is crucial to precisely identify manipulated regions, as intricate details such as subtle artifacts and boundary distinctions take precedence. The localization model, therefore, benefits from architectures well-suited for object identification or segmentation, exemplified by the choice of the ViT architecture. This modular approach acknowledges the divergent demands of classification and localization, tailoring each model to excel in its designated role, allowing for more efficient training and task-specific optimization as required by each architecture used.

4.1.1 ELA-DenseNet

The ELA-DenseNet binary classification model fuses the ELA forensic technique along with a pre-trained DenseNet-121 backbone to determine the authenticity of the query images. The primary aim of the development process was to utilize ELA specifically and then find the optimal CNN architecture that complements it. The model uses a specialized method to apply ELA to the images as a preprocessing step that then returns ELA images in a format similar to Figure ??. These ELA images are given as input to a CNN model. Different CNNs architectures were experimented with, which will be elaborated on in the results section, and the optimal architecture for our use-case was the DenseNet-121. The model was trained on the CASIA V2 dataset for forgery classification.

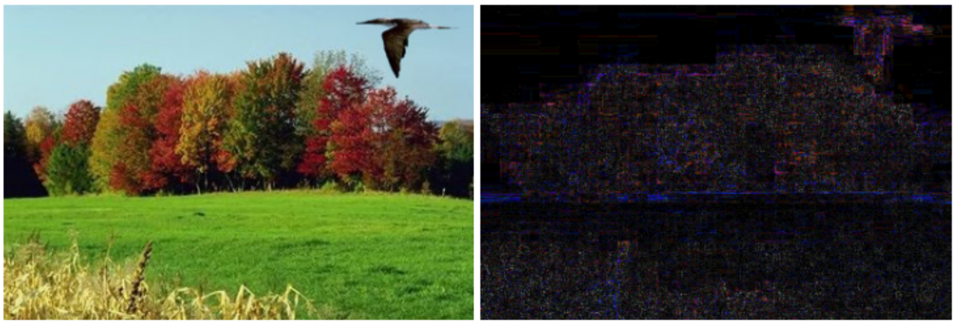

Figure 4.1: The left shows a tampered image and the right is its corresponding ELA image

DenseNet-121 is a CNN architecture that emphasizes dense connectivity between layers. It introduces the concept of dense blocks, where each layer is connected to every other layer in a feed-forward manner [?]. This dense connectivity promotes feature reuse and gradient propagation, addressing issues like vanishing gradients. DenseNet-121 consists of convolutional layers, dense blocks, and transition layers. The convolutional layers perform initial feature extraction, while the dense blocks facilitate information flow by concatenating feature maps from previous layers. Transition layers control the number of feature maps and spatial dimensions. Specifically, DenseNet-121 consists of four dense blocks, three transition layers and a total of 121 layers (117-convolution, 3-transition, and 1-classification). It has been successful in image classification tasks. It achieves high accuracy with fewer parameters compared to other architectures [?]. This parameter efficiency can lead to more compact models and reduced risk of overfitting, especially in situations with limited training data, which is currently the case in image splicing detection.

The tailored model takes a pre-trained DenseNet121 model, unfreezes its layers for fine-tuning, and adds additional layers on top to adapt it for the image splicing detection task. Firstly, it initializes a base model using a pre-trained DenseNet121 model and then loads the pre-trained weights from the ImageNet dataset. The ImageNet dataset is a large visual dataset containing 1,281,167 training images, 50,000 validation images and 100,000 test images. Subsequently, all layers within the base model are set as trainable, enabling fine-tuning during subsequent training. Additional layers are then appended to the base model's output, including a 2D convolutional layer with 1024 filters, a global average pooling layer, a flattening layer, and fully connected dense layers with varied neuron counts and activation functions; ReLU was used throughout the model and softmax was used in the output layer. A dropout layer is incorporated to mitigate overfitting.

To further prevent overfitting and enhance model training, various configurations were implemented. They include a learning rate scheduler, ReduceLROnPlateau, which dynamically adjusts the learning rate during training based on the validation loss. If the loss plateaus for a certain number of epochs (patience), the learning rate is reduced by a factor of 0.2, with a minimum allowable learning rate. Additionally, early stopping is implemented through EarlyStopping, monitoring validation accuracy and terminating training if no improvement is observed within a specified patience period. The chosen optimizer is RMSprop, a variant of stochastic gradient descent (SGD) that adapts the learning rate for each parameter individually, with a learning rate of 0.0005, a rho parameter of 0.9, epsilon set to 1e-08, and no decay. Finally, the model is compiled using this optimizer, employing binary crossentropy as the loss function and accuracy as the metric for model evaluation during training. The model was set to train for 30 epochs but terminated at the 27th epoch due to the EarlyStopping mechanism.

4.1.2 Fine-tuned Vision Transformer

In this proposed method, the aim was to finetune the pretrained IML-ViT [?] transformer and use its foundational architecture to improve the ability of ViTs to generalize to unseen online images. The motivation for choosing this model lied in its relative simplicity and innovative implementation by the authors who outlined three unique points that they wished to tackle in image manipulation localization tasks: high resolution, multi-scale, and edge supervision.

In the context of IML, maintaining high resolution is emphasized to address the information-intensive nature of IML tasks, which focus on intricate artifacts rather than object-level macro-semantics. Existing methods, while utilizing various extractors to trace artifacts, often compromise first-hand artifacts through resizing methods. Therefore, preserving the original image resolution is crucial to retain essential details for the model to effectively learn. Additionally, edge supervision becomes paramount in IML as the primary objective is to distinguish between tampered and authentic regions, with a particular emphasis on the boundary of the tampered region. Unlike typical semantic segmentation tasks that identify information within a target region, visible artifacts in IML are concentrated along the periphery of the manipulated region. Consequently, the model is guided to concentrate on learning the distribution of manipulated region edges for optimal performance. Additionally, addressing the variability in tampered area percentages across different IML datasets, incorporating multi-scale supervision during pre-processing and model design stages is deemed essential. This approach helps enhance generalization across diverse datasets with varying proportions of tampered areas, addressing challenges arising from labor-intensive dataset creation and limited sizes in IML datasets such as CASIA V2 and Columbia.

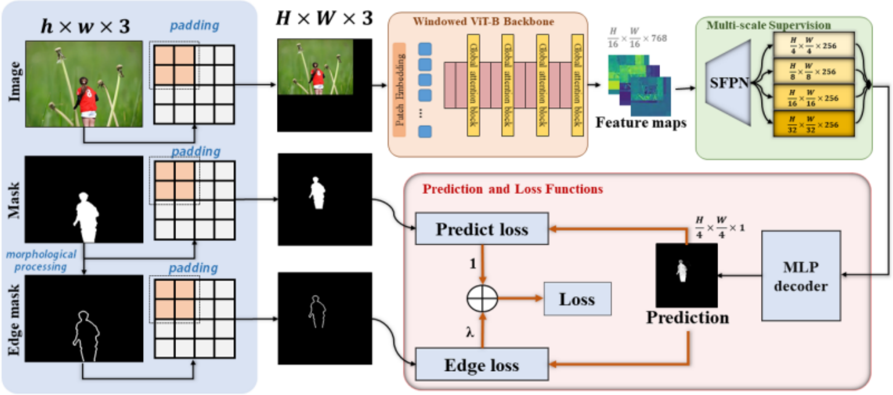

Figure 4.2: Overview of the general structure of IML-ViT [?].

The model starts with a ViT backbone, which applies self-attention and feed-forward layers to extract meaningful features from the input image. This backbone consists of transformer layers that capture global dependencies and relationships between different parts of the image. After the ViT backbone, a Simple Feature Pyramid Network (FPN) is used to generate a multi-scale representation of the features. The FPN combines features from different scales to capture both fine-grained and high-level information from the image. Next, the features from the FPN are passed through an MLP Predict Head, which consists of fully connected layers that predict the manipulated regions of the image. The output of the MLP is a mask representing manipulated regions, where high values indicate the presence of objects and low values indicate the background. The model computes the loss by comparing the predicted mask to the ground truth mask using Binary Cross Entropy (BCE) loss, and it also introduces an edge loss term to encourage accurate manipulation detection along object edges. To generate the manipulated image, the predicted mask is upsampled using bilinear interpolation, and the manipulated regions are obtained by applying the sigmoid function to the predicted mask.

The available source code for the model was pretrained on the CASIA V2 dataset, meaning another dataset had to be used for fine-tuning. The InTheWild dataset [?] was chosen. It includes 201 spliced images curated from online sources such as THE ONION, a parody (fake) news website, and REDDIT PHOTOSHOP BATTLES, an online community of users who create and share manipulated images.

The fine-tuning process involves adapting the model's parameters to the intricacies of the specific dataset, enabling it to learn and discern patterns indicative of splicing forgeries. By utilizing the learned representations from the pretrained IML-ViT, this method aimed to enhance the accuracy and generalization capabilities of the model, especially to images that typical localization models have proven to not perform well on.

4.2 Proposed Dataset

In the pursuit of advancing image splicing localization tasks, a main challenge lies in the scarcity of publicly available and representative datasets. Despite efforts to access datasets like the Fantastic Reality Dataset [?], attempts to acquire relevant data were hindered by unresponsiveness from dataset authors, underscoring a critical gap in the availability of diverse datasets crucial for training and evaluating models in image splicing detection. Currently available datasets that are popularly used in research in the field are listed in Table **??**. InTheWild, the selected dataset for fine-tuning the ViT model, designed for an online-centric use-case, is suitable but constrained by its small size. Recognizing the inadequacy of the existing dataset, there emerged a compelling need to expand upon it. This imperative stemmed from the necessity to augment the dataset's diversity and size, ensuring a more robust foundation for refining models tailored to the challenges of online image manipulation scenarios. To address this, the proposed dataset, an expansion of the InTheWild dataset [?], was curated, providing a larger training dataset for fine-tuning the ViT. The objective was to enhance the ViT's ability to generalize by incorporating a representative and diverse dataset from multiple creators, reflecting samples encountered realistically online.

With the primary goal of expanding the dataset, we sourced the 208 additional images from the same online platforms, namely THE ONION and REDDIT PHOTOSHOP BATTLES. This approach aimed to uphold the dataset's homogeneity. However, with the aim of grater representation, the new images display a greater variety of post-processing and more complex subject matters and even different visual styles. The incorporation of these new images resulted in more than doubling the size of the initial dataset from 201 to 409. Since no groundtruth masks exist for online images, manual annotations were performed, and the corresponding groundtruth masks were created using Labelbox, with a specific emphasis on leveraging their advanced AutoSegment 2.0 feature. The new dataset was labeled ExpandedInTheWild. Figure **??** displays samples along with their groundtruth masks.

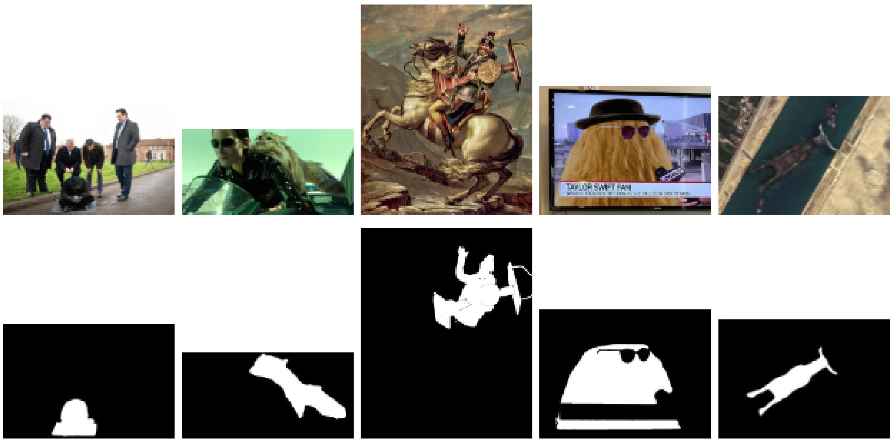

Figure 4.3: Samples from the curated ExpandedInTheWild dataset

Dataset Name	Forgery Type	Real/Forged Images	Image Format	Post-processing
CASIA V1	Splicing, copy move, removal	800/921	TIFF, JPEG, BMP	No
CASIA V2	Splicing, copy move, removal	7491/5123	TIFF, JPEG, BMP	Yes
Columbia Grey	Splicing	933/912	BMP	No
Columbia Color	Splicing	183/180	TIFF	No
Carvalho	Splicing	100/100	PNG	Yes
IMD2020	Splicing, copy move, removal	35000/35000	JPEG	Yes
Dresden	Splicing, copy move, removal	-/16961	JPEG	No
Fantastic Reality	Splicing	12000/1000	JPEG	Yes
Realistic Tampering	Object insertion, removal	220/220	TIFF	Yes
DEFACTO	Splicing, copy move, removal	-/229000	JPEG, PNG, TIFF	Yes
InTheWild	Splicing, copy move, removal	-/201	JPEG, PNG	Yes

Table 4.1: Available image forgery detection datasets

Chapter 5

Experiments & Results

5.1 ELA-DenseNet

In this section, we delve into the experimental journey of the proposed classification model before deciding on the DenseNet-121 architecture as the preferred CNN backbone making up the final ELA-DenseNet model. These preliminary experiments explore various architectures, methodologies, and fine-tuning strategies, providing insights into the model's evolution and guiding the ultimate selection of DenseNet-121. This section presents a retrospective analysis of the experimental iterations, paving the way for a comprehensive understanding of the model's development and the rationale behind its architectural choices.

Initially, the first models implemented were simple sequential CNNs trained from scratch without leveraging any pre-trained weights. These models, as with all subsequent models, were trained on the CASIA V2 dataset. These 'primitive' models generalized very poorly despite configurations to reduce overfitting. This behavior was the motivation for a shift towards transfer learning and adopting pre-trained models as backbone architectures.

After extensive research, the CNN-backbone candidates were chosen, namely: AlexNet, DenseNet-121, Xception and VGG16. They were chosen based on their proven ability to perform well on image classification tasks [?]. All of the models were used in conjunction with the ELA component. Training time was an important factor in the selection process given limited computational resources and the time-constrained nature of the project. All models had the pre-trained ImageNet weights loaded and were set to train for 30 epoch with an RMSprop optimizer and an early stopping mechanism as described in the Proposed Model section above. Notably, the Xception model took almost 4x as much training time as the DenseNet-121 model despite an expected similar training time due to their relatively similar architecture complexity.

The metrics used to evaluate the models included accuracy, recall, and F1-score. Accuracy, a basic measure reflecting overall correctness, is calculated by dividing correct predictions (true positives and true negatives) by the total number of predictions. Precision and recall, key components of the F1-score, assess positive prediction accuracy and the model's ability to capture all positive instances, respectively. The F1 score combines these metrics into a single value, providing a comprehensive assessment of model performance on a scale from 0 to 1. A higher F1 score indicates a better balance between precision and recall. Comparative results for candidate models are presented in Table **??**. The DenseNet model outperformed others, achieving 93.2% accuracy and an F1-Score of 0.881 with a 0.959 recall on the CASIA V2 dataset.

In the realm of image forgery detection, especially in combating misinformation, achieving good recall is paramount. Recall measures the model's effectiveness in identifying all instances of forged images, reducing the risk of misinformation going undetected. High recall, prioritized to capture as many instances of image forgery as possible, helps minimize false negatives and ensures a comprehensive identification of deceptive content. While this emphasis on recall may come at the expense of precision, the overarching objective is to reduce instances of misinformation evading detection. In the battle against misinformation, a robust recall is fundamental for fostering public trust and upholding the integrity of information channels, contributing to a more reliable and secure digital landscape.

$$\begin{aligned}
\text{Accuracy} &= \frac{TP + TN}{TP + TN + FP + FN} \\
\text{Precision} &= \frac{TP}{TP + FP} \\
\text{Recall} &= \frac{TP}{TP + FN} \\
\text{F1 Score} &= \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \times TP}{2 \times TP + FP + FN}
\end{aligned}$$

Base Model	Accuracy	Precision	Recall	F1-Score
AlexNet	0.831	0.850	0.767	0.806
VGG16	0.887	0.802	0.758	0.778
Xception	0.907	**0.874**	0.876	0.875
DenseNet-121	**0.932**	0.856	**0.959**	**0.901**

Table 5.1: Comparisons of the candidate models on CASIA V2

To benchmark against existing ELA-based splicing detection models such as [**?**] and the original ELA model outlined in [**?**], our proposed ELA-DenseNet underwent evaluation using the Columbia dataset. Additionally, the state-of-the-art MiniNet deep learning model [**?**], chosen for its absence of an ELA component, required manual testing on the Columbia dataset due to its initially reported results being based on experiments with the CASIA V2 dataset. The results in Table **??** indicate that ELA-DenseNet surpasses these benchmarks, achieving an accuracy of 95.1% and an F1-Score of 0.936 with emphasis on a higher recall.

Model	Accuracy	Precision	Recall	F1-Score
ELA	0.223	0.796	0.113	0.237
ELA-LBP	0.914	0.901	0.932	0.917
MiniNet	0.867	0.852	0.791	0.819
ELA-DenseNet	**0.951**	**0.913**	**0.966**	**0.936**

Table 5.2: Comparisons of the proposed model against benchmarks on the Columbia dataset

5.2 Fine-tuned Vision Transformer

The performance of the ViT-IML model is thoroughly examined by training it on the InTheWild [**?**] dataset, revealing an initially suboptimal performance of the original model as shown under the ‘OG Predict’ and ‘After thresholding’ masks. Finetuning the model involved initially employing the Adam optimizer, as shown in Figure **??**, and subsequently transitioning to the Stochastic Gradient Descent (SGD) optimizer, as shown in Figure **??**.

To ensure a fair and meaningful comparison, the performance of the models was tested on the CASIA V1 dataset and the Columbia dataset - the same datasets used to derive the published results for the original ViT-IML model. The metrics used for evaluation are pixel-level F1 score with a fixed threshold 0.5 and pixel-level Area Under Curve (AUC), which are commonly used metrics in previous works. In some evaluation methods, determining the optimal threshold for the F1 score can be challenging due to variations in real-world data distributions. To address this, the pixel-level F1 score is proposed as an alternative metric, using a fixed threshold of 0.5. This approach offers a practical and interpretable measure for assessing a model’s performance at the pixel level, making it more suitable in scenarios where finding an optimal threshold is uncertain or impractical. The use of a consistent threshold enhances the applicability and reliability of the evaluation process, especially in real-world, dynamic environments. This evaluation and its metrics aim to determine the overall impact of fine-tuning on the new model’s ability to generalize when compared directly with its predecessor. By conducting these tests on the familiar ground of the CASIA V1 and Columbia datasets, we establish a consistent benchmark to validate improvements and ascertain the model’s efficacy in real-world scenarios.

The results of the tests are shown in Table **??**. The comparative analysis demonstrates that the SGD optimizer yields superior results. The table also shows the results of fine-tuning on the ExpandedInTheWild dataset, more than double the size of the InTheWild. As expected, the model’s performance improved with enriched data with an F1 score of 0.721 and AUC of 0.915 on the CASIA V1 dataset compared to an F1 score of 0.704 an AUC of 0.892 from fine-tuning on the smaller dataset.

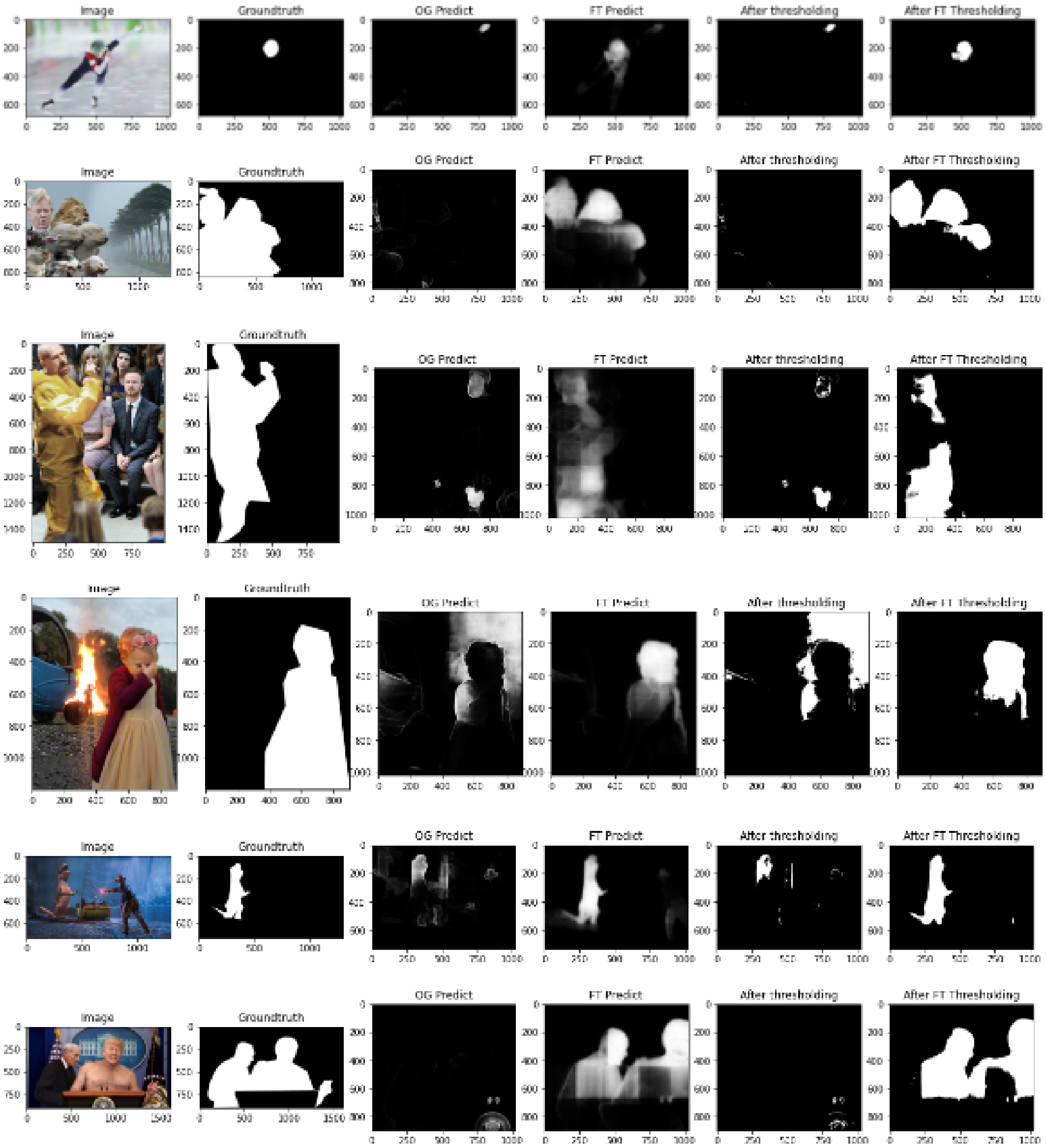

Figure 5.1: Comparison of before and after fine-tuning (FT) with Adam

Model	CASIA V1		Columbia	
	PL F1	PL AUC	PL F1	PL AUC
ViT-IML	*0.658*	*0.867*	*0.836*	*0.908*
Adam-ViT-IML	0.673	0.874	0.842	0.912
SGD-ViT-IML	**0.704**	**0.892**	**0.864**	**0.922**
SGD-ViT-IML (Expanded)	***0.721***	***0.915***	***0.887***	***0.931***

Table 5.3: Results of the fine-tuned models on test datasets. The bolded italics show the results after fine-tuning on the ExpandedInTheWild dataset.

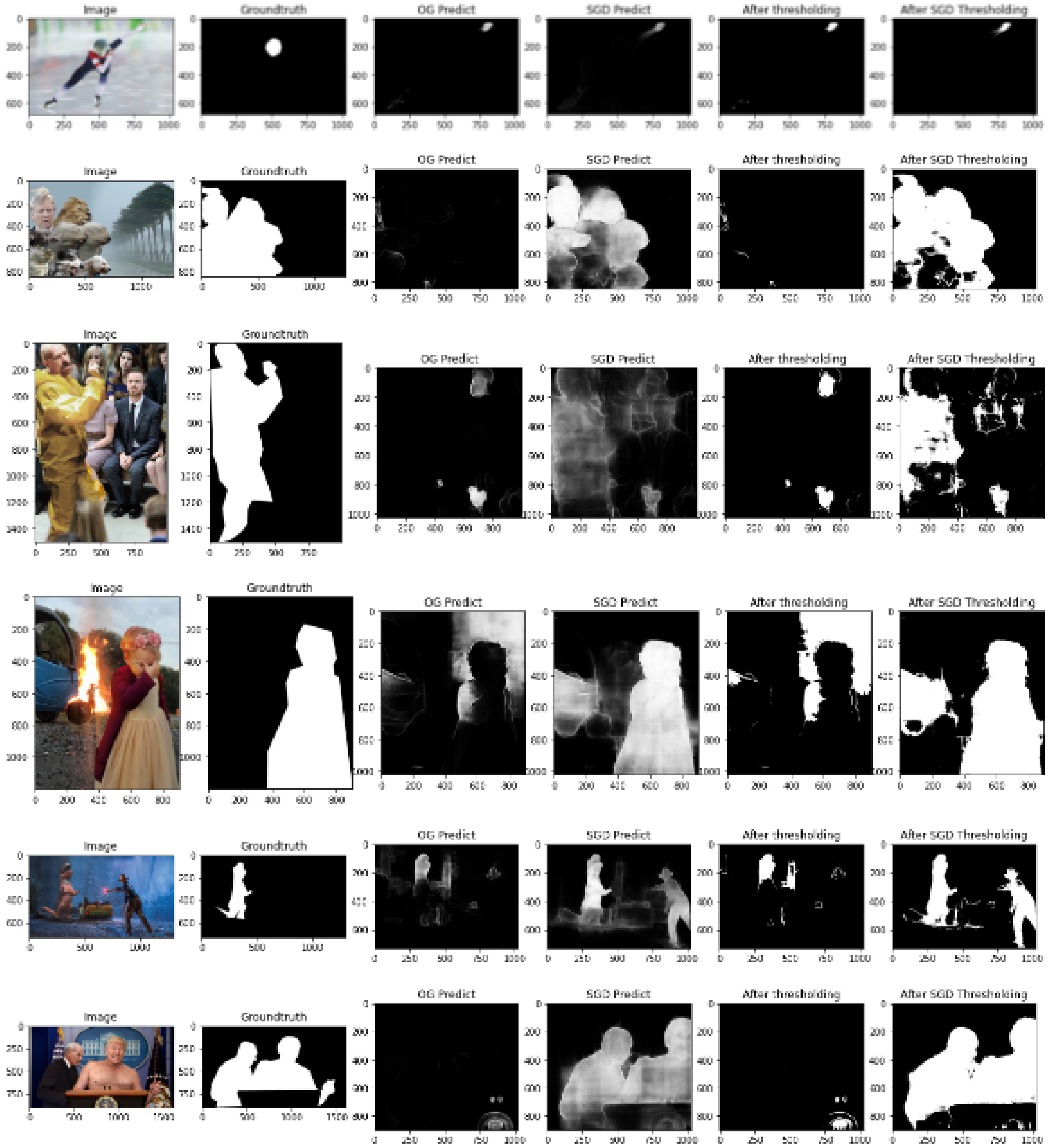

Figure 5.2: Comparison of before and after fine-tuning with SGD

Chapter 6

Conclusion

In conclusion, this thesis addresses the pressing issue of image forgery and its impact on the authenticity and integrity of visual information. By focusing on image splicing detection, the thesis provides insights into the classification and localization of manipulated areas within images. The proposed models, the ELA-DenseNet and the fine-tuned vision transformer, demonstrate promising results in detecting image splicing, contributing to the field of digital forensics. Additionally, the enriched dataset enhanced the generalization and performance of the localization model and provides good representative data for forgeries encountered in realistic scenarios online. The findings of this thesis highlight the importance of robust algorithms and quality datasets in combating image forgery-induced disinformation. By preserving trust and integrity in visual information, the proposed methods contribute to the mitigation of the negative consequences of image forgery in various contexts, including security, legal, and social domains.

This thesis opens up several avenues for future research in the field of image forgery detection. Firstly, the proposed models can be further optimized and fine-tuned to improve their performance and generalizability. Exploring different network architectures, incorporating additional features, or utilizing advanced optimization techniques could potentially enhance the accuracy and robustness of the models. Secondly, the research can be extended to develop real-time forgery detection systems that can process images in a streaming fashion, enabling the detection of manipulated content in near real-time scenarios. Finally, as image forgery techniques continue to evolve, it is crucial to continuously update and expand the dataset used for evaluation to ensure the effectiveness of the proposed models in detecting new and sophisticated forms of forgery. By addressing these future directions, researchers can further advance the field of image forgery detection and contribute to the development of reliable and robust techniques for preserving the integrity of visual information.

Appendix

List of Figures

List of Tables

Bibliography

[1] Belal Ahmed, T Aaron Gulliver, and Saif alZahir. Image splicing detection using mask-rcnn. *Signal, Image and Video Processing*, 14:1035–1042, 2020.

[2] Shams Forruque Ahmed, Md. Sakib Alam, Maruf Hassan, Mahtabin Rodela Rozbu, Taoseef Ishtiak, Nazifa Rafa, M. Mofijur, A. B. Shawkat Ali, and Amir H. Gandomi. Deep learning modelling techniques: Current progress, applications, advantages, and challenges. *Artificial Intelligence Review*, 56(11):13521–13617, 2023.

[3] Amani Alahmadi, Muhammad Hussain, Hatim Aboalsamh, Ghulam Muhammad, George Bebis, and Hassan Mathkour. Passive detection of image forgery using dct and local binary pattern. *Signal, Image and Video Processing*, 11:81–88, 2017.

[4] Md Zahangir Alom, Tarek M Taha, Christopher Yakopcic, Stefan Westberg, Paheding Sidike, Mst Shamima Nasrin, Brian C Van Esesn, Abdul A S Awwal, and Vijayan K Asari. The history began from alexnet: A comprehensive survey on deep learning approaches. *arXiv preprint arXiv:1803.01164*, 2018.

[5] Khurshid Asghar, Xianfang Sun, Paul L Rosin, Mubbashar Saddique, Muhammad Hussain, and Zulfiqar Habib. Edge–texture feature-based image forgery detection with cross-dataset evaluation. *Machine Vision and Applications*, 30(7-8):1243–1262, 2019.

[6] Chaitra Basavaraj and P. Reddy. Digital image forgery: taxonomy, techniques, and tools–a comprehensive study. *International Journal of System Assurance Engineering and Management*, 14, 12 2022.

[7] Hippolyte Bayard. Portrait as a drowned man, 10 1840.

[8] Dulari Bhatt, Chirag Patel, Hardik Talsania, Jigar Patel, Rasmika Vaghela, Sharnil Pandya, Kirit Modi, and Hemant Ghayvat. Cnn variants for computer vision: History, architecture, application, challenges and future scope. *Electronics*, 10(20):2470, 2021.

[9] Tiziano Bianchi, Alessia De Rosa, and Alessandro Piva. Improved dct coefficient analysis for forgery localization in jpeg images. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2444–2447. IEEE, 2011.

[10] Jason Bunk, Jawadul H Bappy, Tajuddin Manhar Mohammed, Lakshmanan Nataraj, Arjuna Flenner, BS Manjunath, Shivkumar Chandrasekaran, Amit K Roy-Chowdhury, and Lawrence Peterson. Detection and localization of image forgeries using resampling features and deep learning. In *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, pages 1881–1889. IEEE, 2017.

[11] Oswald Campesato. *Artificial intelligence, machine learning, and deep learning.* Mercury Learning and Information, 2020.

[12] Deirdre Carmody. Time responds to criticism over simpson cover. *New York Times*, page 8–8, Jun 1994.

[13] Beijing Chen, Xiaoming Qi, Yiting Wang, Yuhui Zheng, Hiuk Jae Shim, and Yun-Qing Shi. An improved splicing localization method by fully convolutional networks. *IEEE Access*, 6:69472–69480, 2018.

[14] Dan Claudiu Ciresan, Ueli Meier, Jonathan Masci, Luca Maria Gambardella, and Jürgen Schmidhuber. Flexible, high performance convolutional neural networks for image classification. In *Twenty-second international joint conference on artificial intelligence.* Citeseer, 2011.

[15] Samantha Cola. We are truly fucked: Everyone is making ai-generated fake porn now. *Vice*, Jan 2018.

[16] Peter Corcoran, Cosmin Stan, Corneliu Florea, Mihai Ciuc, and Petronel Bigioi. Digital beauty: The good, the bad, and the (not-so) ugly. *Consumer Electronics Magazine, IEEE*, 3:55–62, 10 2014.

[17] Seana Coulson. *Semantic leaps: Frame-shifting and conceptual blending in meaning construction.* Cambridge University Press, 2001.

[18] Davide Cozzolino and Luisa Verdoliva. Single-image splicing localization through autoencoder-based anomaly detection. pages 1–6, 12 2016.

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[20] Mohammed Fakhrulddin Abdulqader, Adnan Yousif Dawod, and Ann Zeki Ablahd. Detection of tamper forgery image in security digital mage. *Measurement: Sensors*, 27:100746, 2023.

[21] Don Fallis. The epistemic threat of deepfakes. *Philosophy amp;amp; Technology*, 34(4):623–643, 2020.

[22] William D. Ferreira, Cristiane B.R. Ferreira, Gelson da Cruz Júnior, and Fabrizzio Soares. A review of digital image forensics. *Computers Electrical Engineering*, 85:106685, 2020.

[23] Jessica Fridrich and Jan Kodovsky. Rich models for steganalysis of digital images. *IEEE Transactions on information Forensics and Security*, 7(3):868–882, 2012.

[24] Helmut Gernsheim. A concise history of photography. Dover, 1986.

[25] Amin Ghiasi, Hamid Kazemi, Eitan Borgnia, Steven Reich, Manli Shu, Micah Goldblum, Andrew Gordon Wilson, and Tom Goldstein. What do vision transformers learn? a visual exploration, 2022.

[26] Teddy Surya Gunawan, Siti Amalina Mohammad Hanafiah, Mira Kartiwi, Nanang Ismail, Nor Farahidah Za'bah, and Anis Nurashikin Nordin. Development of photo forensics algorithm by detecting photoshop manipulation using error level analysis. *Indonesian Journal of Electrical Engineering and Computer Science*, 7(1):131–137, 2017.

[27] Xiao Guo, Xiaohong Liu, Zhiyuan Ren, Steven Grosz, Iacopo Masi, and Xiaoming Liu. Hierarchical fine-grained image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3155–3165, 2023.

[28] A. Guterres. Countering disinformation for the promotion and protection of human rights and fundamental freedoms, Aug 2022.

[29] Jing Hao, Zhixin Zhang, Shicai Yang, Di Xie, and Shiliang Pu. Transforensics: Image forgery localization with dense self-attention, 2021.

[30] Robert M Haralick, Karthikeyan Shanmugam, and Its' Hak Dinstein. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6):610–621, 1973.

[31] J. Henley. 85% of people worry about online disinformation, global survey finds, Nov 2023.

[32] János Horváth, Sriram Baireddy, Hanxiang Hao, Daniel Mas Montserrat, and Edward J. Delp. Manipulation detection in satellite images using vision transformer, 2021.

[33] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017.

[34] Zhaoyang Huang, Han Zhou, Yijin Li, Bangbang Yang, Yan Xu, Xiaowei Zhou, Hujun Bao, Guofeng Zhang, and Hongsheng Li. Vs-net: Voting with segmentation for visual localization, 2021.

[35] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7014–7023, 2018.

[36] Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A Efros. Fighting fake news: Image splice detection via learned self-consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 101–117, 2018.

[37] Khawar Islam. Recent advances in vision transformer: A survey and outlook of recent work, 2023.

[38] Vincent Itier, Olivier Strauss, Laurent Morel, and William Puech. Color noise correlation-based splicing detection for image forensics. *Multimedia Tools and Applications*, 80:13215–13233, 2021.

[39] Daniel Jeronymo, Yuri Campbell, and Leandro Coelho. Image forgery detection by semi-automatic wavelet soft-thresholding with error level analysis. *Expert Systems with Applications*, 85, 05 2017.

[40] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012.

[41] N. Jurrat. Mapping digital media: Citizen journalism and the internet, july 2011.

[42] Navdeep Kanwal, Akshay Girdhar, Lakhwinder Kaur, and Jaskaran S Bhullar. Digital image splicing detection technique using optimal threshold based local ternary pattern. *Multimedia Tools and Applications*, 79(19-20):12829–12846, 2020.

[43] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022.

[44] Mohammed S. Khazaal, Monji Kherallah, and Faiza Charfi. An overview on detecting digital image splicing. In *2022 International Arab Conference on Information Technology (ACIT)*, pages 1–4, 2022.

[45] Jan Kietzmann, Linda W Lee, Ian P McCarthy, and Tim C Kietzmann. Deepfakes: Trick or treat? *Business Horizons*, 63(2):135–146, 2020.

[46] Vladimir Vladimirovich Kniaz, Vladimir Alexandrovich Knyaz, and Fabio Remondino. The point where reality meets fantasy: Mixed adversarial generators for image splice detection. In *Neural Information Processing Systems*, 2019.

[47] Neal Krawetz. A picture's worth... hacker factor solutions, 2007.

[48] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[49] Yingxin Lai, Zhiming Luo, and Zitong Yu. Detect any deepfakes: Segment anything meets face forgery detection and localization, 2023.

[50] Hugo Larochelle, Yoshua Bengio, Jérôme Louradour, and Pascal Lamblin. Exploring strategies for training deep neural networks. *Journal of machine learning research*, 10(1), 2009.

[51] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[52] Shiguo Lian and Yan Zhang. *Multimedia Forensics for Detecting Forgeries*, pages 809–828. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[53] Zhouchen Lin, Junfeng He, Xiaoou Tang, and Chi-Keung Tang. Fast, automatic and fine-grained tampered jpeg image detection via dct coefficient analysis. *Pattern Recognition*, 42(11):2492–2501, 2009.

[54] Grace W Lindsay. Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of cognitive neuroscience*, 33(10):2017–2031, 2021.

[55] Bo Liu and Chi-Man Pun. Locating splicing forgery by fully convolutional networks and conditional random field. *Signal Processing: Image Communication*, 66:103–112, 2018.

[56] Yahui Liu, Enver Sangineto, Wei Bi, Nicu Sebe, Bruno Lepri, and Marco Nadai. Efficient training of visual transformers with small datasets. *Advances in Neural Information Processing Systems*, 34:23818–23830, 2021.

[57] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[58] Weiqi Luo, Jiwu Huang, and Guoping Qiu. Jpeg error analysis and its applications to digital image forensics. *IEEE Transactions on Information Forensics and Security*, 5(3):480–491, 2010.

[59] Jiayi Ma, Xingyu Jiang, Aoxiang Fan, Junjun Jiang, and Junchi Yan. Image matching from handcrafted to deep features: A survey. *International Journal of Computer Vision*, 129(1):23–79, 2020.

[60] Xiaochen Ma, Bo Du, Zhuohang Jiang, Ahmed Y. Al Hammadi, and Jizhe Zhou. Iml-vit: Benchmarking image manipulation localization by vision transformer, 2023.

[61] DMA Mahawatta and L Ranathunga. Image splice detection through noise pattern analysis. In *2018 4th International Conference for Convergence in Technology (I2CT)*, pages 1–6. IEEE, 2018.

[62] Mohammad Shahjahan Majib, Md Mahbubur Rahman, TM Shahriar Sazzad, Nafiz Imtiaz Khan, and Samrat Kumar Dey. Vgg-scnet: A vgg net-based deep learning framework for brain tumor detection on mri images. *IEEE Access*, 9:116942–116952, 2021.

[63] Zane Mathews. Fun or fear: Deepfake app puts celebrity faces in your selfies. *Vice*, Mar 2020.

[64] Kunj Bihari Meena and Vipin Tyagi. *Image Forgery Detection: Survey and Future Directions*, pages 163–194. 04 2019.

[65] Kunj Bihari Meena and Vipin Tyagi. Image splicing forgery detection techniques: A review. In Mayank Singh, Vipin Tyagi, P. K. Gupta, Jan Flusser, Tuncer Ören, and V. R. Sonawane, editors, *Advances in Computing and Data Sciences*, pages 364–388, Cham, 2021. Springer International Publishing.

[66] Kunj Bihari Meena and Vipin Tyagi. Image splicing forgery detection using noise level estimation. *Multimedia Tools and Applications*, pages 1–18, 2023.

[67] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization, 2022.

[68] Minati Mishra and Flt Adhikary. Digital image tamper detection techniques-a comprehensive study. *arXiv preprint arXiv:1306.6737*, 2013.

[69] Sadiq Muhammed T and Saji K. Mathew. The disaster of misinformation: A review of research in social media. *International Journal of Data Science and Analytics*, 13(4):271–285, 2022.

[70] Nicolas M Müller, Karla Pizzi, and Jennifer Williams. Human perception of audio deepfakes. In *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*, pages 85–91, 2022.

[71] S. Murugesan. Understanding web 2.0. *IT Professional*, 9(4):34–41, Aug 2007.

[72] V Parameswaran Nampoothiri and N Sugitha. Digital image forgery—a threaten to digital forensics. In *2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*, pages 1–6. IEEE, 2016.

[73] Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers, 2021.

[74] Michael A Nielsen. *Neural networks and deep learning*, volume 25. Determination press San Francisco, CA, USA, 2015.

[75] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015.

[76] Daniel Palmer. The rhetoric of the jpeg. In *The Photographic Image in Digital Culture*, pages 149–164. Routledge, 2013.

[77] Aditya Pandey and Anshuman Mitra. Detecting and localizing copy-move and image-splicing forgery. *arXiv preprint arXiv:2202.04069*, 2022.

[78] Azra Parveen, Zishan Khan, and Syednaseem Ahmad. Block-based copy–move image forgery detection using dct. *Iran Journal of Computer Science*, 2, 06 2019.

[79] Micheal Peres. page 24–25. Focal Press, 4 edition, 2007.

[80] Ana Pérez-Escoda, Luis Miguel Pedrero-Esteban, Juana Rubio-Romero, and Carlos Jiménez-Narros. Fake news reaching young people on social networks: Distrust challenging media literacy. *Publications*, 9(2):24, 2021.

[81] Nam Thanh Pham, Jong-Weon Lee, Goo-Rak Kwon, and Chun-Su Park. Efficient image splicing detection algorithm based on markov features. *Multimedia Tools and Applications*, 78:12405–12419, 2019.

[82] Thales Pomari, Guillherme Ruppert, Edmar Rezende, Anderson Rocha, and Tiago Carvalho. Image splicing detection through illumination inconsistencies and deep learning. In *2018 25th IEEE International conference on image processing (ICIP)*, pages 3788–3792. IEEE, 2018.

[83] Yao Qiang, Chengyin Li, Prashant Khanduri, and Dongxiao Zhu. Interpretability-aware vision transformer, 2023.

[84] E Ramadhani. Photo splicing detection using error level analysis and laplacian-edge detection plugin on gimp. *Journal of Physics: Conference Series*, 1193:012013, 04 2019.

[85] Yuan Rao and Jiangqun Ni. A deep learning approach to detection of splicing and copy-move forgeries in images. pages 1–6, 12 2016.

[86] Yuan Rao, Jiangqun Ni, and Huimin Zhao. Deep learning local descriptor for image splicing detection and localization. *IEEE Access*, 8:25611–25625, 2020.

[87] Judith Redi, Wiem Taktak, and Jean-Luc Dugelay. Digital image forensics: A booklet for beginners. *Multimedia Tools Appl.*, 51:133–162, 10 2011.

[88] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[89] Paolo Rota, Enver Sangineto, Valentina Conotter, and Christopher Pramerdorfer. Bad teacher or unruly student: Can deep learning say something in image forensics analysis? In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 2503–2508. IEEE, 2016.

[90] Bo-Kai Ruan, Hong-Han Shuai, and Wen-Huang Cheng. Vision transformers: State of the art and research challenges, 2022.

[91] Jesus Ruiz-Santaquiteria, Gloria Bueno, Oscar Deniz, Noelia Vallez, and Gabriel Cristobal. Semantic versus instance segmentation in microscopic algae detection. *Engineering Applications of Artificial Intelligence*, 87:103271, 2020.

[92] Ruslan Salakhutdinov and Geoffrey Hinton. Deep boltzmann machines. In David van Dyk and Max Welling, editors, *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 448–455, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 16–18 Apr 2009. PMLR.

[93] Jitendra Sharma and Rohita Sharma. Analysis of key photo manipulation cases and their impact on photography. 2018.

[94] Surbhi Sharma and Umesh Ghanekar. Spliced image classification and tampered region localization using local directional pattern. *International Journal of Image, Graphics and Signal Processing*, 11:35–42, 03 2019.

[95] Vikas Srivastava and Sanjay Yadav. Texture operator based digital image splicing detection using eltp technique. pages 345–348, 12 2020.

[96] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7262–7272, 2021.

[97] Xiu Su, Shan You, Jiyang Xie, Mingkai Zheng, Fei Wang, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. Vitas: Vision transformer architecture search, 2021.

[98] Sasha Targ, Diogo Almeida, and Kevin Lyman. Resnet in resnet: Generalizing residual architectures. *arXiv preprint arXiv:1603.08029*, 2016.

[99] Hans Thisanke, Chamli Deshan, Kavindu Chamith, Sachith Seneviratne, Rajith Vidanaarachchi, and Damayanthi Herath. Semantic segmentation using vision transformers: A survey. *Engineering Applications of Artificial Intelligence*, 126:106669, 2023.

[100] Shobhit Tyagi and Divakar Yadav. Mininet: a concise cnn for image forgery detection. *Evolving Systems*, 14(3):545–556, 2023.

[101] Hafiz ur Rhhman, Muhammad Arif, Anwar Ullah, Sadam Al-Azani, Valentina Emilia Balas, Oana Geman, Muhammad Jalal Khan, and Umar Islam. Comparative analysis of various image splicing algorithms. In *Soft Computing Applications: Proceedings of the 8th International Workshop Soft Computing Applications (SOFA 2018), Vol. II 8*, pages 211–228. Springer, 2021.

[102] Vinorth Varatharasan, Hyo-Sang Shin, Antonios Tsourdos, and Nick Colosimo. Improving learning effectiveness for object detection and classification in cluttered backgrounds. pages 78–85, 11 2019.

[103] Rohit Verma and Jahid Ali. A comparative study of various types of image noise and efficient noise removal techniques. *International Journal of advanced research in computer science and software engineering*, 3(10), 2013.

[104] James Vincent. This app uses neural networks to put a smile on anybody's face. *The Verge*, Jan 2017.

[105] Gregory K Wallace. The jpeg still picture compression standard. *IEEE transactions on consumer electronics*, 38(1):xviii–xxxiv, 1992.

[106] Jiaqi Wang, Zhengliang Liu, Lin Zhao, Zihao Wu, Chong Ma, Sigang Yu, Haixing Dai, Qiushi Yang, Yiheng Liu, Songyao Zhang, Enze Shi, Yi Pan, Tuo Zhang, Dajiang Zhu, Xiang Li, Xi Jiang, Bao Ge, Yixuan Yuan, Dinggang Shen, Tianming Liu, and Shu Zhang. Review of large vision models and visual prompt engineering. *Meta-Radiology*, page 100047, 2023.

[107] Junke Wang, Zuxuan Wu, Jingjing Chen, Xintong Han, Abhinav Shrivastava, Ser-Nam Lim, and Yu-Gang Jiang. Objectformer for image manipulation detection and localization, 2022.

[108] Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity, 2020.

[109] Xinyi Wang, He Wang, Shaozhang Niu, Jiwei Zhang, et al. Detection and localization of image forgeries using improved mask regional convolutional neural network. *Mathematical Biosciences and Engineering*, 16(5):4581–4593, 2019.

[110] Baptiste Wicht. *Deep Learning feature Extraction for Image Processing.* PhD thesis, 01 2018.

[111] Esra Odabaş Yildirim and Güzin Ulutaş. Image splicing detection with dwt domain extended markov features. In *2018 26th Signal Processing and Communications Applications Conference (SIU)*, pages 1–4. IEEE, 2018.

[112] Esra ODABAŞ YILDIRIM and Güzin ULUTAŞ. Markov-based image splicing detection in the dct high frequency region. In *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*, pages 1–4. IEEE, 2018.

[113] Mohammed Zakariah, Muhammad Khurram Khan, and Hafiz Malik. Digital multimedia audio forensics: past, present and future. *Multimedia tools and applications*, 77:1009–1040, 2018.

[114] Marcello Zanardelli, Fabrizio Guerrini, Riccardo Leonardi, and Nicola Adami. Image forgery detection: a survey of recent deep-learning approaches. *Multimedia Tools and Applications*, 82(12):17521–17566, 2023.

[115] Wei Zhai, Pingyu Wu, Kai Zhu, Yang Cao, Feng Wu, and Zheng-Jun Zha. Background activation suppression for weakly supervised object localization and semantic segmentation, 2023.

[116] Lifei Zhan and Yuesheng Zhu. Passive forensics for image splicing based on pca noise estimation. In *2015 10th International Conference for Internet Technology and Secured Transactions (ICITST)*, pages 78–83. IEEE, 2015.

[117] Y Zhang, T Shi, and Zhe-Ming Lu. Image splicing detection scheme based on error level analysis and local binary pattern. *Netw. Intell*, 6:303–312, 2021.