Preface

The decision to focus on the field of Digital Image Forensics for my thesis was driven by my personal interest in online deception, specifically the uncovering it, and a strong desire to contribute something of value to a field that is becoming increasingly important in our digital society. My educational background positioned me to comprehend the complexities of this discipline and build technical solutions.

Throughout my thesis, I encountered numerous challenges, all of which afforded me invaluable experience and learning opportunities. This thesis is the result of extensive literature review, technical development, experimentation and iterative refinement of all these aspects.

I wish to express my sincere gratitude to everyone who played pivotal roles in this journey. First and foremost, I would like to thank my Promotor, Prof. Dr. Lode Jorrisen, for granting me the opportunity to embark on this project and for allowing me the freedom to explore and pursue my interests. His flexibility enabled me to add a large degree of personal touch to this work, while still guiding me toward the creation of a valuable contribution to the field.

I also want to thank my Co-Promotor Steven Moonen, whose consistent guidance, expertise, feedback, and insights were instrumental in steering this work in the right direction and elevating it to a higher standard. His contributions were central to the development of this thesis.

I also thank Bram Vanherle for his participation in numerous meetings and for providing valuable feedback and insights that were important to the progress of this work.

My appreciation also goes to Prof. Dr. Kris Aerts and Kathleen Bovin for their support, they helped me keep up morale and focus at key moments.

Finally, I want to thank my friends and family for their unconditional support which gave me strength to carry on, even through difficult times.

I hope this work will deliver a valuable contribution to the field of Digital Image Forensics and that the open-source tools I developed will prove useful to others. I also hope this work inspires readers to further contribute to this field.

This work was made possible with support from MAXVR-INFRA, a scalable and flexible infrastructure that facilitates the transition to digital-physical work environments. This project is subsidized by the Flemish Government and the European Union.

Table of Contents

List of Tables

List of Figures

Terminology

- **Adversarial Attack**: A process of deliberately modifying input data to cause an AI network to misclassify or produce incorrect outputs. These attacks exploit vulnerabilities in AI models, often with the intent of causing errors or misleading the model.
- **Artificial Intelligence (AI)**: A branch of computer science focused on creating systems capable of performing tasks that typically require human intelligence.
- **Area Under the Curve (AUC)**: A measure of the overall performance of a classification model, represented as the area under the ROC curve. An AUC of 0,5 is equivalent to random guessing. According to Nahm [35], an AUC of 0.7 to 0.8 is considered acceptable, 0.8 to 0.9 is considered excellent, and more than 0.9 is considered outstanding. In practice, however, the required classification performance of a model is determined by its application purpose.
- **Fast Fourier Transform (FFT)**: An efficient algorithm to compute the Discrete Fourier Transform (DFT) and its inverse.
- **Fourier Domain**: A mathematical space in which signals or images are represented in terms of their frequencies, rather than their spatial or temporal values.
- **Fourier Transform**: A mathematical transform that converts a function or signal from its original domain (often time or space) into a representation in the frequency domain.
- **False Positive Rate (FPR)**: A measure that indicates the proportion of actual negatives that are incorrectly identified as positives by a model. It represents the likelihood of a model incorrectly classifying a non-event as an event.
- **Generative Adversarial Network (GAN)**: An AI network consisting of two components: a generator that creates fake data and a discriminator that attempts to distinguish between real and fake data. Both components are trained simultaneously, with the generator aiming to fool the discriminator, and the discriminator improving its accuracy.
- **JPEG Compression**: A widely used lossy compression algorithm that reduces the file size of images by discarding some of the image data, leading to a reduction in quality. The compression process is irreversible, meaning the lost information cannot be recovered.
- **JPEG Quality**: JPEG quality refers to the level of compression applied to an image. For example, a JPEG quality of 100% retains the original image detail, but results in larger file size. A lower quality, e.g. 60%, reduces the file's size at the expense of image quality.
- **Oversaturation**: A phenomenon in digital images where colors are rendered with excessive intensity, often leading to a loss of detail. Oversaturation can occur due to post-processing, image compression, or deliberate editing.
- **Receiver Operating Characteristic Curve (ROC curve)**: A graph that plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. The ROC curve provides insight into the performance of a classification model across different thresholds.
- **True Positive Rate (TPR)**: A measure that indicates the proportion of actual positives that are correctly identified by a model. It represents the model's sensitivity or ability to detect true events.

Abstract

With the rise of realistic AI-generated images and continuously advancing photo-editing software, it has become increasingly difficult to reliably distinguish between authentic and manipulated images. Using Digital Image Forensics, the primary objective of this work is to conduct a comprehensive quantitative and qualitative study that compares traditional forensic techniques to a state-of-the-art AI based approach. A practical underpinning for this work is the development of a toolset capable of providing reliable and transparent evidence regarding the authenticity of an image.

Results indicate that the developed algorithms were implemented correctly. The quantitative comparison suggests that the combined traditional techniques outperform a recent state-of-the-art AI network by an average of 17% more detections in scenarios where 0% False Positives are allowed. However, the study acknowledges potential biases in the validation process and further experimentation is necessary to ascertain the reliability of these findings. A qualitative comparison suggests that traditional techniques are more dependable than AI.

This work emphasizes the importance of developing reliable digital image forensic tools and outlines a future vision where AI can be utilized in a key support role to assist forensic analysts. Guided by an open-source philosophy, each algorithm was successfully integrated into Sherloq, an established open source image forensic toolset, which has garnered positive feedback from the community.

Abstract in het Nederlands

Met de opkomst van realistische beelden gegenereerd door AI en de evolutie van beeldverwerkingssoftware wordt het steeds moeilijker om onderscheid te maken tussen authentieke en gemanipuleerde foto's.

In dit werk wordt, met behulp van Digitale Beeldforensica, een uitgebreide kwantitatieve en kwalitatieve studie uitgevoerd waarin traditionele technieken worden vergeleken met een state-of-the-art AI-netwerk voor de detectie van gemanipuleerde beelden. Een praktische insteek voor dit werk is de ontwikkeling van een reeks tools die op een betrouwbare manier bewijs kunnen leveren over de authenticiteit van een foto.

De resultaten suggereren dat de ontwikkelde algoritmen correct zijn geïmplementeerd. De kwantitatieve studie laat zien dat de traditionele technieken gecombineerd gemiddeld 17% meer detecties opleveren dan een recent state-of-the-art AI-netwerk in scenario's waarin 0% vals-positieven zijn toegestaan. De studie erkent ook mogelijke tekortkomingen in het validatieproces; verdere experimenten zijn nodig om deze bevindingen te bevestigen. De kwalitatieve vergelijking suggereert dat traditionele technieken betrouwbaarder zijn dan een AI.

Dit werk benadrukt het belang van het ontwikkelen van betrouwbare tools voor forensisch onderzoek van digitale beelden en schetst een visie waarin AI gebruikt kan worden in een ondersteunende rol voor forensische analisten. Geleid door een open-source filosofie is elk algoritme geïntegreerd in Sherloq, een bestaand open source beeldforensische toolset, waarop positieve reactie gekomen is.

1. Introduction

"Seeing is believing" is a well-known phrase that is losing its meaning on the internet. In the past, creating a doctored image required specialized knowledge and skill. However, as programs like Adobe Photoshop have been developed and increased their functionalities, the skill required to make believable fakes has decreased over time. With the advent of AI, an individual only needs a computer and an internet connection to create the most beautiful and photo realistic images using programs such as stable diffusion, DALL-E, Midjourney, inpainting, and many others.

An age is emerging where "seeing is believing" is no longer true. How are people supposed to filter fact from fiction in the digital world? That's where the field of Digital Image Forensics comes in.

1.1 Current state of online deception

GANs are becoming increasingly more realistic and while we enjoy this technology for creative and entertainment purposes, the misuse of this technology is becoming a growing concern [16].

A recent study [1] details studies on how humans are increasingly unable to differentiate between real and computer generated faces. The average human accuracy for identifying GAN-faces was found to be around 50-60% [17]. Another peculiar finding was that synthetic faces were rated as appearing more trustworthy than real faces, by an average of 7%. Similar findings were replicated in other studies [3][4].

Similar advancements have been made in other areas, such as audio and video manipulation[1]. A concerning development is called "puppet master" technology, which is advancing from head to full body synthesis. This deepfake technology allows a computer system to learn the characteristics of a person based on a picture or video input. Anyone can subsequently use this system to imitate the original person, thus becoming a puppet master. It's worth noting that the advancement of this technology might not necessarily be driven by the desire to deceive, but rather by a well-meaning intent of creating full-body and realistic avatars for virtual reality applications [7][8].

For audio, the advancements are equally alarming. In 2019, Forbes and The Wall Street Journal reported on a cybercrime where criminals used AI generated voice technology to impersonate the CEO of a company, tricking an employee into transferring $243,000 to a fraudulent account[5][6].

Combining the advances in video, audio and text, it's no surprise that AI influencers have emerged on social media, with varying degrees of realism. These influencers sometimes have millions of followers and business models are being developed around them [9][10][11].

What's more, thousands and possibly millions of people all around the world are already forming intimate relationships with AI companions [66] and it is an understatement that the infamous Nigerian prince scam [65] has long been eclipsed by this new wave of deception. As the line between real and fake becomes ever smaller, it raises a worrying question: How many people are unknowingly following highly realistic AI accounts and are forming emotional bonds with them?

2. Literature Study

2.1 General principles in Digital image Forensics

When navigating the field of digital image forensics [26,28,59], the following principles are important to keep in mind.

1. The field of Digital image forensics is a contest between better forging methods and detection methods. All detection techniques can in theory be circumvented by a sophisticated forger.
2. There exists no such thing as proof of authenticity. There is only a lack of conclusive evidence of tampering.
3. The field of image forensics is evidence based. Real photographs can appear fake, and fake images can appear real. Tampering can be proven, but does not necessarily entail that the contents of the image are 'not truthful'. For example, a contrast enhanced image can still represent a real person, location or situation.

2.2 Current state of Artificial Intelligence in digital image forensics

Research in Digital Image Forensics has increasingly focused on the development of AI methods to detect manipulated images and AI generated images. This shift is understandable, because AI generally outperforms individual traditional detection methods. For example, detecting exact copies of natural objects, such as trees or clouds, in an image can serve as strong evidence of tampering. Although the development of effective copy detection algorithms is a research topic in its own right, well-trained AI models have been shown to outperform traditional approaches in both accuracy and speed [67]. Another reason for focus on AI can be attributed to AI's versatility, which is one of its key advantages over traditional image forensics. It can be applied to various situations, provided that the necessary training data is available.

There are downsides to consider when using AI in digital image forensics. Multiple survey studies [18, 19 , 20, 21, 22] come to similar conclusions:

- **Reliance on Specific Cues or input format:** AI detection models often depend on particular cues or input formats, leading to narrow applicability and vulnerability to adversarial attacks.
- **Data Dependency & data availability:** AI models often require access to large amounts of data. The lack of publicly available data hinders the ability of others to replicate, evaluate, and improve proposed models. Data may not be shared due to privacy concerns, legal restrictions, infrastructure limitations or competition between researchers.
- **Model attribution**: AI remains largely a "black box," meaning that its decision-making process is not transparent. For legal accountability, it is crucial to provide explanations beyond statistical models based on big data.
- **Robustness and generalization**: Many AI models perform well on training data, but degrade to chance performance on unseen data or fail to retain accuracy when the original data is degraded For example, by JPEG compression.

Among the various applications of AI in image forensics, the detection of deepfakes has received particular attention. The urgency of identifying deepfakes, especially GAN-generated faces, has driven extensive research in this area. A recent survey by XinWang et al. [16] focused on the

detection of GAN-generated faces and divides research methods in 4 categories: (1) deep learning-based methods, (2) physical-based methods, (3) physiological-based methods, and (4) human visual performance. We will briefly surmise the findings for each of these categories to broaden the understanding of AI driven research in the field.

- **(1) Deep Learning-Based Methods**: These approaches have demonstrated impressive overall performance. The biggest downside to this approach according to XinWang et al. is the model attribution (black box). The real world favors explainability as a basis of trust.
- **(2) Physical-Based Methods**: These methods were found to be more robust to adversarial attacks and their results are more explainable and intuitive. A common drawback in this category was the need for specific circumstances to obtain a reliable result. For example, some methods required the input to be a high quality frontal portrait of a face for reliable detections.
- **(3) Physiological-Based Methods**: These approaches focus on investigating specific features of the human face. Many of these methods have become obsolete due to advancements in GANs. For example, early GANs generated faces with mismatched eye-pupil colors due to a lack of understanding of human anatomy. As GAN networks become more advanced, some these obvious fake artifacts have started to disappear. An artifact that is still detectable in modern GAN networks is the shape of the pupil [16,12], which can have irregular shapes and is supposed to be circular or ellipse shaped. It is inevitable that this technique will eventually become obsolete as GANs improve. However, that does not subtract from the current effectiveness of this detection method, nor from its potential use as a tool for detection and analysis. After all, not every forger has the same level of expertise to create undetectable fakes.
- **(4) Human Visual Performance**: Human visual performance is based on detecting physiological cues from generated images (e.g. pupil shapes). Human performance is largely biased and studies that were conducted come to similar conclusions that humans are already unable to distinguish between real and GAN-generated faces.

XinWang et al. concluded that while GAN-face detection has made notable progress, there is still significant room for improvement.

2.3 Recent studies & their most important findings

A recent study by Gonzalo et al. [46] focused intentionally on the narrow task of distinguishing real faces from fake ones in order to combat the creation of fake online media profiles.

Their model, at a false positive rate (FPR) of 0.5%, achieved a true positive rate (TPR) of 98.0% when tested on images synthesized by generators that were included in the training dataset (Evaluation Set A). When tested on images synthesized by unseen generators (Evaluation set B), the model's TPR dropped by 13,5% to 84.5% at the same FPR. Remarkably, the model detected 0% of generated images when no face was present in the image, suggesting that it has identified a face-specific artifact common to most synthetic faces.

The model's robustness was also tested under various conditions, retaining a TPR of 88% at a JPEG quality of 60% and a TPR of 91% at a 128x128 resolution, both at an FPR of 0.5%. These results further suggest that the model has latched onto a general-purpose artifact and not a low-level artifact.

When compared with other state-of-the-art methods [47,48], the researchers demonstrated the superiority of their model. The model presented by Corvi et al. [47], which exploited Fourier

artifacts, reported an AUC of 90% in their own testing. However, it achieved only a 23.8% TPR at a 0.5% FPR on Evaluation Set A. This significant drop in performance highlights two interesting phenomena: Firstly, testing results may be biased, leading to an overestimation of a model's capabilities in the original work. Secondly, a recent state of the art model, fails to generalize to new data one year later. The authors hypothesize that this discrepancy in performance is due to the more challenging dataset used in their work.

Another model created by Mundra et al. [48] focused on GAN-generated faces and achieved a 99.5% TPR at a 1% FPR on evaluation set A. Its performance drops to 86% TPR (1% FPR) for evaluation set B. These performances are similar at a slightly higher FPR. However, this model fails to detect faces generated by diffusion models.

Upon analyzing the training datasets used in these studies, it is plausible that the observed differences in performance could largely be attributed to the quantity and diversity of the training data. The notion that AI models perform better when trained on more and higher quality training data is a consistent finding for AI research [51,52,53].

Dong et al. [72] in their work highlight the vulnerability of large neural network-based detection strategies to adversarial attacks. They reviewed fake image detectors that analyze the frequency spectrum of GAN-generated images, because these approaches have demonstrated superior performance in recent works. Dong et al. proposed a few methods capable of imperceptibly altering images that would subsequently go undetected. Their work demonstrated that detection accuracy from some networks dropped to a mere 50% after processing GAN images that were initially detected at accuracy rates of over 90%. These findings underscore the need for more robust methods that are capable of reliably identifying GAN-based forgeries.

2.3 Conclusion

From the extensive body of research reviewed, several key insights have emerged. AI-driven approaches demonstrate superior accuracy in detecting image manipulations compared to classical techniques. However, these AI-driven methods also present significant challenges that are difficult to overcome, particularly their heavy reliance on the quality and quantity of training data. These limitations persist even when narrowly focusing on specific tasks, such as distinguishing AI-generated faces from real ones. The task of distinguishing real faces from synthetic ones has seen significant development since at least 2018 [49,50]. This observation suggests that an end is not yet in sight and further highlights the dynamic arms race between detection methods and increasingly sophisticated methods of generating fakes.

Taking this arms race to its logical conclusion suggests that AI-generated fakes will eventually become indistinguishable from real images, making them effectively undetectable. When this future inevitably arrives, alternative solutions must be developed to determine what is real and what is not.

Given this context, there are three possible approaches for researching Digital Image Forensics, namely:

1. **Competing in the Arms Race:**
    - **Developing New Detection Methods:** This involves creating a new technique that tries to exploit information in a novel way to detect manipulations or develop a new AI model.

- **Focusing on Adversarial Attacks:** Another approach could be to create more sophisticated fakes, thereby introducing new challenges for the community to solve.

2. **Exploring Alternative Solutions for Combating Fake Images:**
    - **Blockchain Technology:** This technology has the potential to create immutable records of digital files, which could be registered as originals. While promising, blockchain solutions are still in their infancy and face significant challenges related to scalability, energy consumption, and public acceptance .
    - **Adobe's Content Authenticity Initiative:** Adobe's initiative aims to attach proof of authenticity to images created using Photoshop. This approach could become a standard in the industry, helping to distinguish between original and manipulated content. However, its effectiveness depends on widespread adoption across different platforms and industries .
    - **Watermarking Technologies:** Embedding digital watermarks into images could serve as a "fingerprint" to trace the origins of content and verify its authenticity. This method, however, may be vulnerable to removal or manipulation, especially as AI continues to evolve .
3. **Focusing on Practical, Here-and-Now Solutions:**
    - There are many problems in the world that require solutions that could be addressed by Digital Image Forensics such as curtailing the spread of misinformation by flagging manipulated images on social media, detecting fake social media profiles and the authentication of images in legal proceedings and criminal investigations.

In this thesis we will focus on developing a practical solution for a specific, current problem. This decision was influenced by several factors:

Limitations of Competing in the Arms Race: As a newcomer, the probability of successfully contributing a groundbreaking technique or providing critical insights is low. The field is highly competitive, and the search space for new solutions is vast, making it likely that our efforts would replicate existing work. Established professionals with years of experience are better equipped to navigate these challenges and innovate effectively.

Challenges of Alternative Solutions: Focussing on authentication of real images rather than detection of fakes, is a forward-looking approach that can potentially solve the root problem. Yet, it may still take many years before this type of solution becomes necessary and our personal interests align more with the third option. Furthermore, big players in the tech industry are already focussing on this topic.

Although AI is a hot topic with significant economic potential, several factors led us to steer clear of this approach:

- **Dependence on Data:** AI models require vast amounts of high-quality data, which is often inaccessible to newcomers. Publicly available datasets are commonly used, but relying on them limits the uniqueness of any new approach and often results in models with poor generalizability when tested on unseen data .
- **Vulnerability to Adversarial Attacks:** AI models are particularly susceptible to adversarial attacks, where slight alterations in the input can lead to incorrect outputs. This vulnerability

undermines the reliability of AI-driven solutions, especially in high-stakes scenarios such as legal proceedings .

- **Black-Box Nature of AI Models:** One of the most significant challenges with AI is its "black box" nature—decisions are made without clear explanations of the underlying processes. While research is ongoing to make AI models more explainable, current methods are not yet sufficiently viable for use in critical decision-making processes .
- **Trust in AI Decisions:** Trusting AI to make important decisions, such as determining the authenticity of an image, is risky. Real-world examples have already shown the pitfalls of over-reliance on AI for making important decisions [56,75].

Given these factors, we decided to focus on classical techniques, which are more transparent and explainable. Our interest veered to legal cases and the ability to provide transparent evidence towards the authenticity of an image or lack thereof.

To our knowledge, no study has attempted to compare a combination of classical techniques to a state-of-the-art AI model. Therefore this work will focus on classical techniques that provide visual outputs capable of localizing manipulations in an image, and compare these to a suitable AI network that pursues the same objective.

3. Method

3.1 Introduction

In our search for a state-of-the-art AI model, we found Multi-Modal Fusion (MM-Fusion) by Triaridis et al. [23], publicly available on GitHub. Their work focuses on localization of manipulations in images and offers two variants: early fusion and late fusion, with the early fusion model slightly outperforming late fusion, achieving an average Area Under Curve (AUC) of 0,897 across five datasets. In our work we used the early fusion model. Because MM-Fusion has a state of the art performance for localizing image manipulations, as demonstrated in their comparative study with other approaches, it was an ideal candidate to utilize in our work.

To diversify our toolkit detection methods, we focused on three different artifacts: JPEG compression, resampling artifacts and noise artifacts. In total we developed three python implementations based on their relevant papers and also developed a user interface for each algorithm. Subsequently we submitted our work to an open source project on Digital Image Forensic, called Sherloq. All our contributions were novel to this project and were accepted by the community.

The rest of this section will introduce Sherloq and discuss the three techniques we contributed to this project, followed by a quantitative and qualitative comparison between MM-Fusion and the three traditional techniques combined.

3.2 Sherloq & Existing Digital Image Forensics tools

Before starting, it is important to note that there are relatively few free projects available in the field of Digital Image Forensics, despite its critical importance. This scarcity is likely due to the specialized knowledge required to develop and effectively use these tools, as well as the economic value these tools and knowledge hold in investigation scenes and legal cases.

Several commercial solutions have emerged [57,58] offering digital images forensic toolkits. These companies often sell their software packages alongside training courses, ensuring that their tools are used effectively by professionals. For example, Cognitec [58] offers its software for a monthly fee of $175, highlighting the economic premium of forensic tools. Amped Software [58] caters exclusively to government agencies or certified professionals.

In contrast to these commercial solutions, Sherloq [63] stands out as an open-source image forensic toolset. Launched in 2015 by Guido Bartoli, a software development engineer with a passion for photography, Sherloq was founded on the belief that security by obscurity is not the best approach for digital image forensics. Bartoli advocates for an open-source philosophy, believing that everyone should be able to try out techniques and verify how they work by looking at the source code. An informed community that shares ideas and knowledge will lead to the development of trustworthy tools. We wholeheartedly support this philosophy.

Since 2020, Sherloq was ported into Python with a graphical interface and as of today some 32 features are offered for analyzing images, three of these features were added by our team this year.

For getting started in the field of digital images forensics, we recommend Neal Krawetz's whitepaper [64]. This work offers thoughtful analyses of doctored images and explains how to effectively use a few techniques. Most of these techniques are available in Sherloq.

To our knowledge, Sherloq is one of a kind. Some other interesting free tools are FotoForensics [59], Forensically [60], Ghiro [61] and Media Verification Assistant [62]. Some of these tools are largely abandoned and limited in scope, yet their websites are still operational and can serve as a starting point for familiarizing oneself in digital image forensics.

3.3 Implemented techniques for the localization of manipulations

In this section, we will provide an overview of each algorithm we implemented and offer a brief technical explanation to facilitate understanding of how this technique detects manipulations. We will demonstrate the correct functionality of these algorithms by reviewing examples that also serve as practical guides for practitioners. Additionally, we will showcase the user interface integrated into Sherloq. All implementation code is available on GitHub[1].

3.3.1 JPEG Ghosts

We will first discuss the concept of JPEG ghosts, which will help understanding the JPEG ghost maps technique.

JPEG ghosts is a technique pioneered by Hany Farid [25] and is based on the idea that different JPEG qualities can be detected in an image.

To understand JPEG ghosts, it's essential to first examine how JPEG compression affects an image. Consider an image initially saved at a quality of 80% and then resaved at 100%. Despite the final image being reported as quality 100, the effective quality is still 80%. Differences between the effective quality and a reported quality can be uncovered and this process is explained in [25]. Sherloq offers a "Quality estimation" tool that implements this technique.

Figure 1 shows the quality estimation curve of an image originally recorded at 80% quality and then resaved at 90% JPEG quality. The curve demonstrates how the average error increases until it matches the reported quality of 90%, at which point the error difference is 0. The error then increases again until it dips to a local minimum at 80%—the original quality—after which the error steadily increases to 100%.

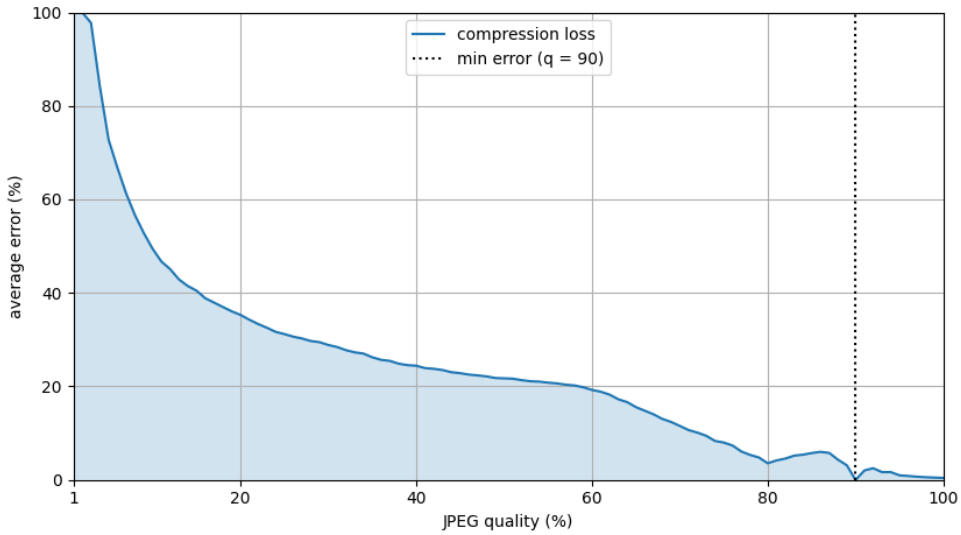

Figure 1: Quality estimation curve for an image originally recorded at 80% quality that was subsequently resaved at 90% JPEG quality. The curve demonstrates how the average error increases until it reaches the reported quality of 90%. Then

[1] *All code available at: https://github.com/UHstudent/digital_image_forensics_thesis*

again at quality 80% the curve reaches a local minima, called a JPEG ghost. From the local minima, the average error steadily increases again.

Figure 2, shows the quality estimation curve when the same image from figure 1 is resaved at 60% quality and then once again at its original quality of 80%. The quality estimation curve shows a local minimum near 60%, which contrasts with the original degradation curve where the average error only increased after surpassing the effective quality of 80%. These local minima that the image reaches after passing its reported quality correspond to the effective quality of the image. The effective quality (60%) that reveals itself when probed is dubbed "the JPEG ghost" by Hany Farid.

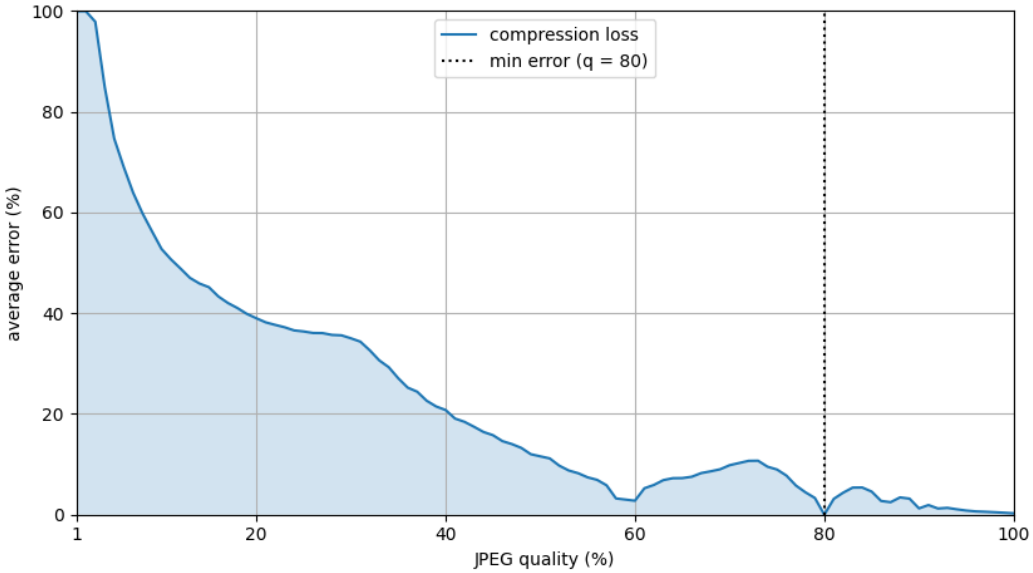

Figure 2: Quality estimation curve for an original image initially saved at 80% quality, which then underwent multiple resaves: first to 90%, then to 60%, and finally back to 80%. The curve shows a local minimum near 60%, indicative of a JPEG ghost. Additionally, the estimation curve becomes somewhat erratic before reaching the quality of 80%.

The presence of JPEG ghosts is proof that an image has been resaved at a different quality than initially recorded. This fact can be used as evidence that an image could have been modified after its initial recording. However there are limitations to this analysis:

1. **Proof of Resaving Only:** JPEG ghost analysis only proves that an image was resaved after its initial recording. It gives no clue about the type of manipulation that has taken place.
2. **Assumption of Similar Compression Algorithms:** The analysis works on the assumption that the JPEG compression settings used by your software and the compression software of the forger are similar. Camera manufacturers and software programs may use varying JPEG compression algorithms. While most JPEG compression operates on an 8x8 grid, some use non-standard grids like 12x12, which could affect analysis [68].
3. **JPEG Block Boundaries:** This analysis only works if the JPEG block boundaries are preserved [68].

3.3.2 JPEG Ghost maps

To localize JPEG ghosts inside an image, the difference between the pixel values can be calculated. The process is thoroughly described by Hany Farid in his paper and book [25, 26]. In this work, we implemented these scientific principles in Python and will demonstrate the correct operation of the algorithm. Additionally, we will qualitatively discuss this technique and offer insights on how practitioners can use it to localize manipulations. A user interface has also been developed and

together with the algorithm has been integrated into Sherloq, making this technique available to practitioners worldwide.

Different cameras and photo editing software employ slightly different JPEG quality scales and quantization tables. To account for this, the image differences are averaged over a block size of 16. As a result, small differences in the original and quantization tables by the algorithm will likely not have a significant impact [25]. A side effect of this approach is that the edges of the image are discarded in the output. Specifically, when the image size is not a multiple of the grid size, any remaining pixels are discarded. For instance, a 501x501 image will produce a ghost map output of 31x31 pixels, meaning that 5 pixels (= 501 – 31 * 16) are discarded near the right and bottom edges of the image. The origin for this process is the top-left corner.

Consider an image of a cat originally captured at quality 80. A central region (100x100) has been extracted, saved at quality 60, and reinserted back into the image at the same location. Figure 3 shows the ghost maps output of this image from JPEG quality 40 to 90 in steps of 5.

At quality 60, the average pixel difference of the central region is near 0, making this region appear black. This black region represents the JPEG ghost, correctly suggestion that a region with an effective JPEG quality of 60 is present in the image.

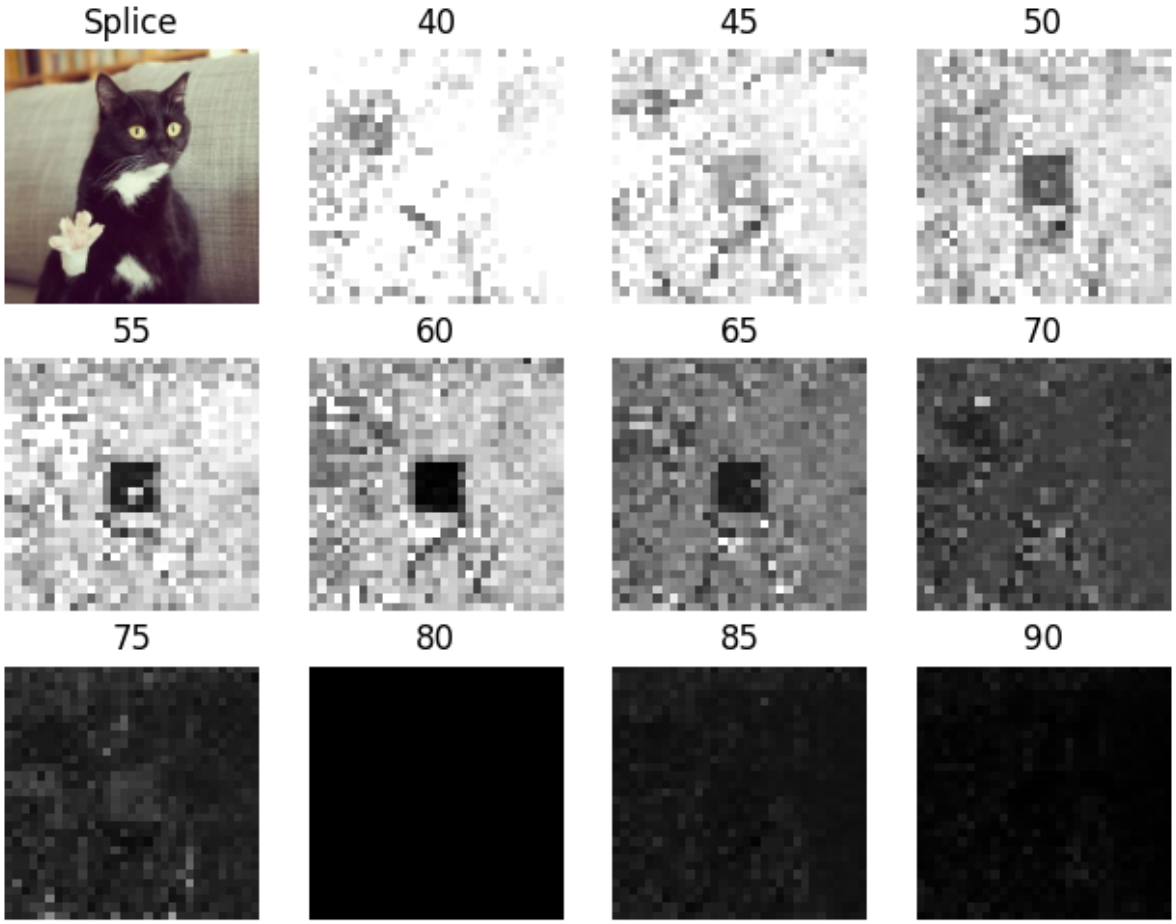

Figure 3: Output maps of the ghost algorithm where a central region of an image, originally saved at JPEG quality 80, is extracted, JPEG compressed at quality 60, and reinserted at the same location. The JPEG ghost is highly salient as a black region at quality 60, correctly indicating the presence of the re-saved region at lower quality.

In the previous example, the manipulated region of quality 60 was inserted in exactly the same place, keeping it aligned with the original JPEG 8x8 grid. When the JPEG grid is not correctly aligned, the JPEG ghost is destroyed. Consider the following example, where a 100x100 pixel region is taken 100 pixels to the right of the image center and then inserted as quality 60 in the center of the original image. This displacement destroys the JPEG ghost as shown in figure 4a.

However, this ghost can be recovered by shifting the image over its X and Y axis for every possible grid alignment (8x8 JPEG grid). This operation guarantees that the grid of the manipulated region will properly align with the JPEG compression grid in one of the 64 offset cases. In this example, the correct offset can be predicted: since the inserted region was taken from 100 pixels to the right, and assuming a standard x and y axis, the correct offset will be x = 100 – 96 (12 * 8 JPEG blocks) = 4 and y = 0. Figure 4b demonstrates that the JPEG Ghost becomes visible at this offset.

It is important to exercise caution when interpreting a ghost plot such as the one in figure 4b. The contrast difference is not as pronounced as a JPEG ghost where the JPEG lattice has not been destroyed. This can be compared to figure 3, where the central region appears as solid black. Only when the JPEG ghost is highly salient it can be confidently identified as an anomaly and thus proof of tampering. When the contrast difference is more subtle and appears as shades of gray, there is a risk of misidentifying a region and producing false positives. In order to confidently identify JPEG ghosts, they must be part of a larger observation: the ghost is a pronounced local minima when compared to its offsets. Figure 4 (a,b & c) demonstrates this local minima phenomenon: the ghost reveals itself at the correct offset (x = 4, y = 0), while the other regions in the image barely change. Thus, a JPEG ghost only reveals itself when properly probed.

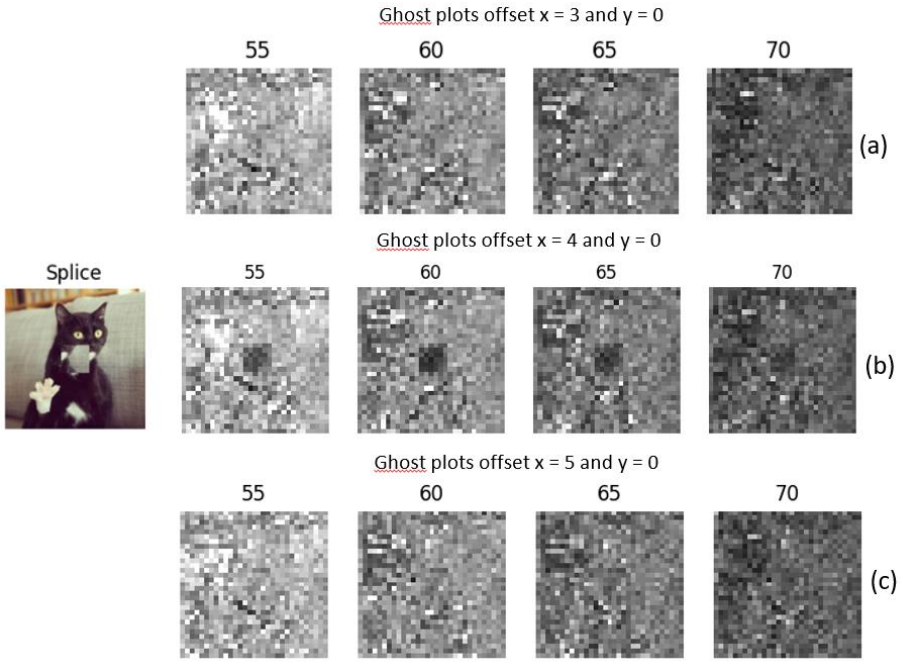

Figure 4: Only the key ghost maps are shown to demonstrate the concepts, in practice these ghost maps were constructed for a larger JPEG quality range, similar to the example in figure 3.
(a,b,c) Example where a region 100 pixels to the right of the central portion has been taken, recompressed at quality 60 and inserted into the central portion of the image.
(b,c) When the JPEG grid is misaligned, the JPEG ghost is destroyed.
(b) When the original grid is properly re-aligned through shifting the image, the JPEG ghost can be recovered.
(a,b,c) Demonstration that a JPEG ghost will only reveal itself when properly probed at the correct JPEG grid alignment. A pronounced local minima can be observed at quality 60 for the correct offset (x = 4, y = 0) compared to its neighboring offsets x = 3 and x = 5. Only when such a pronounced local minima is identified, can JPEG ghosts be confidently identified.

The original paper by Hany Farid [25] concludes that a disadvantage of the JPEG ghost method is “that it is only effective when the tampered region is of lower quality than the image into which it was inserted”. While this is accurate, we believe that this statement can lead to confusion, because JPEG ghosts can also be effective at detecting tampered regions of higher quality than the image into which it was inserted.

To better encapsulate the method’s potential and avoid confusion, the following precise statement should always be used when reasoning about JPEG ghosts: as Farid states: "This approach explicitly detects if part of an image was compressed at a lower quality than the saved JPEG quality of the entire image."

To explore the idea that JPEG ghosts can also be effective at detecting tampered regions of higher quality than the image into which it was inserted, we will retake the first example shown in figure 3 with a twist. A central region of the image is compressed at quality 60, reinserted into the original image with quality 80, and then the entire image is resaved at quality 90. Figure 5 demonstrates that the JPEG ghost at quality 60 is still visible, but at quality 80 a gray box has appeared around the manipulated area. Why?

This can be explained by the 8x8 JPEG grid. The central region (100x100) that was taken and reinserted is not a perfect multiple of an 8x8 JPEG grid. Consequently, at the edges of this region, composite grid blocks are created. These blocks consist of a few pixels from a 60% quality JPEG block and a few pixels from an 80% JPEG block. These composite blocks are then recompressed into new 90% JPEG blocks, which is the gray box that is observed at quality 80 in figure 5.

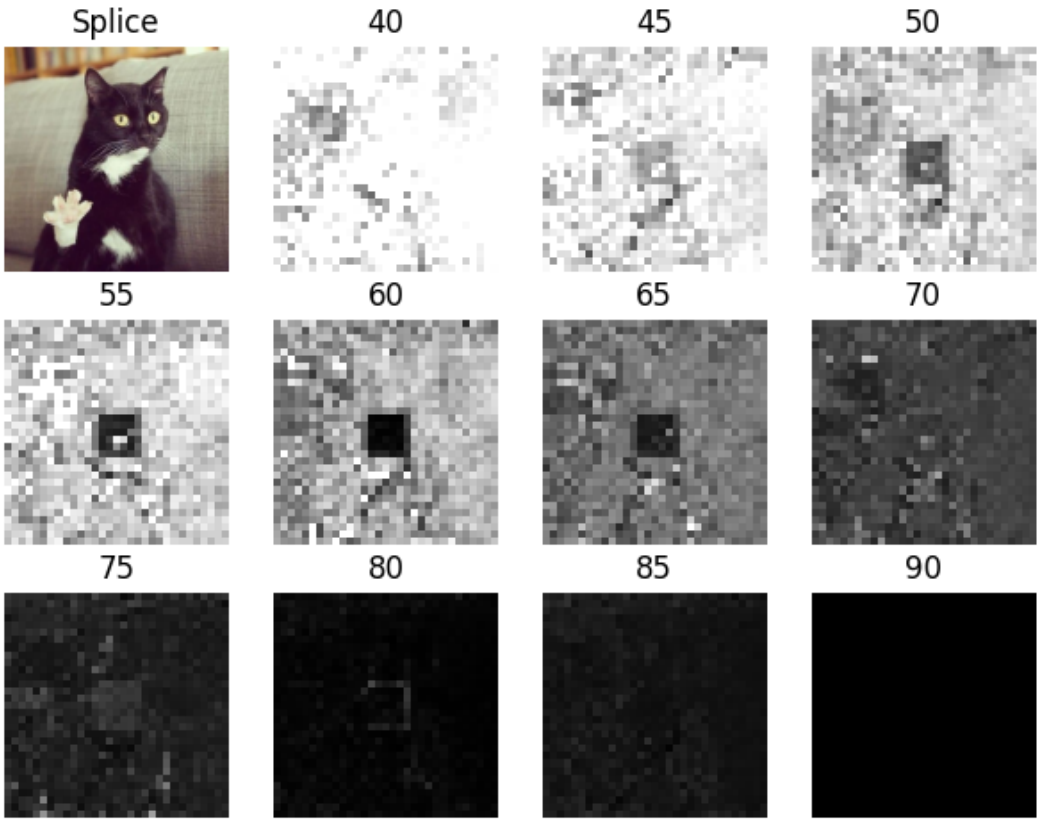

Figure 5: Illustration of a JPEG ghost at quality 60 and 80, where a (100x100) central region compressed at quality 60 was reinserted into the original image and then the entire image was resaved at quality 90. A gray box appears around the manipulated area at quality 80 due to the creation of composite JPEG blocks because a 100x100 region is not a multiple of the original 8x8 JPEG grid.

To illustrate this idea more clearly, consider a scenario where a part compressed at 60% quality, located 150 pixels to the left of the central portion (x-offset = -150), is inserted into a picture of quality 80, and the result is saved at quality 90. Figure 6a shows the result of this operation.

At quality 80 the following is detected: the effective quality of the image is 80, thus it becomes highly salient at its original quality level. The misaligned JPEG grid in the center causes these blocks to become fresh 90% JPEG blocks. This phenomenon would not appear only if a forger copies and pastes a region exactly from and to the same 8x8 JPEG grid, a low chance. Additionally, if the region is not a multiple of the JPEG grid, composite blocks will be created that risk detection.

The central JPEG ghost of quality 60 can be recovered when shifting the image to the correct 8x8 JPEG grid alignment. Knowing the original offset was x = -150, the correct offset to recover the inserted 60% region can be predicted: x =-150 + 144 (18*8) = -6; thus the correct grid offset is x = -6 or x = 2. Figure 6b shows the result of this operation. Remember that this JPEG ghost can only be reliably identified because it represents a pronounced local minimum when analyzing the grid offsets.

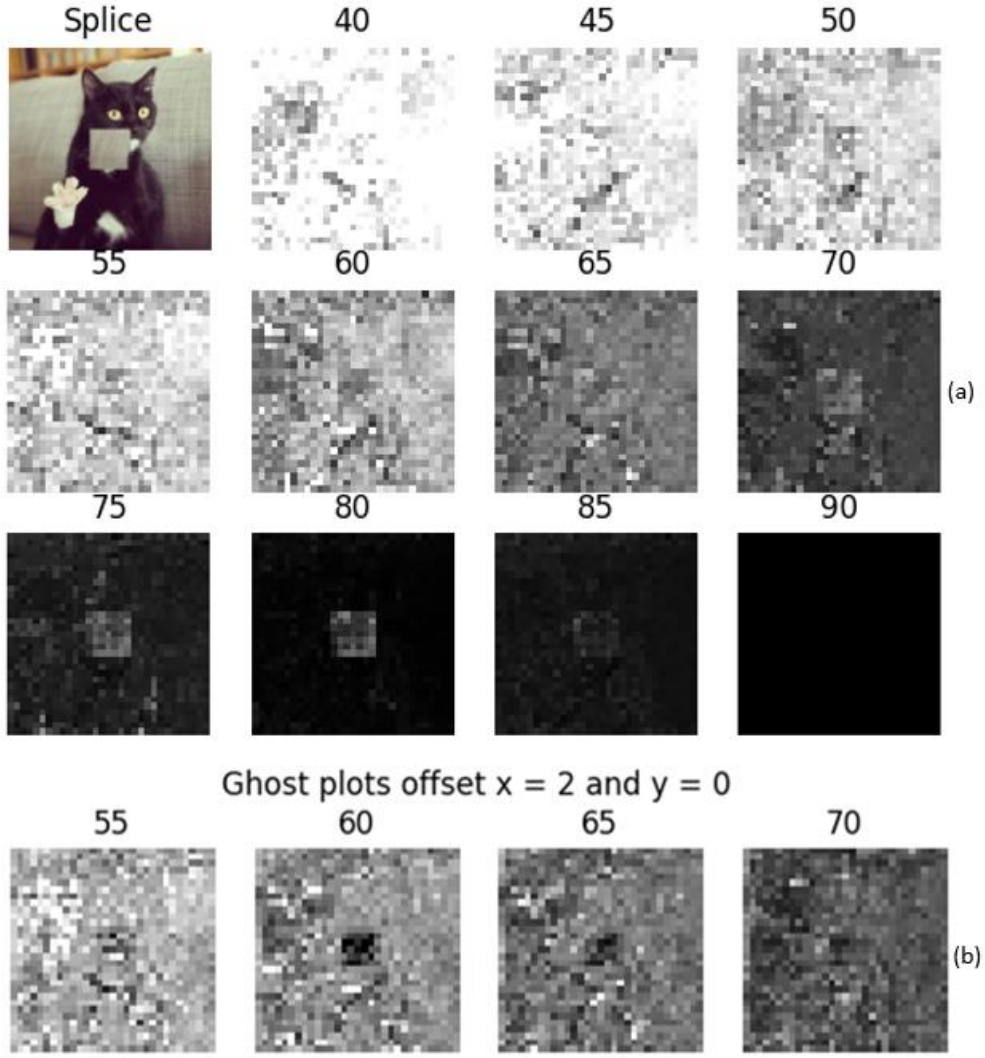

Figure 6: (a) Illustration of a JPEG ghost at quality 80, where a (100x100) region 150 pixels to the right of the central portion has been taken, recompressed at quality 60 and re-inserted into the central portion of the image, which then was resaved at quality 90. Because the JPEG grid from the inserted 60% JPEG quality region was misaligned, new JPEG blocks of quality 90 were created. The JPEG ghost that is detected is not the manipulated region in this case, it is the inserted region. (b) Key ghost maps at offset (x = 2, y = 0), revealing the presence of the original JPEG ghost at quality 60.

In conclusion, changing the content of an 8x8 JPEG grid becomes highly salient when the modified image is saved at a higher quality than the original image. Figures 7 and 8 illustrate this idea with practical examples.

In figure 7 the image of a cat has been modified. Because the manipulated image was saved at a higher quality than the original image, the manipulations become highly salient: A new paw, the operation glove around the original claw and a doctor's head mirror. The ghost maps show that the majority of the image is of a lower quality than these added objects, a JPEG ghost. It is because the cat and the rest of the image are presumed to be authentic, that the added objects are identified to be manipulations.

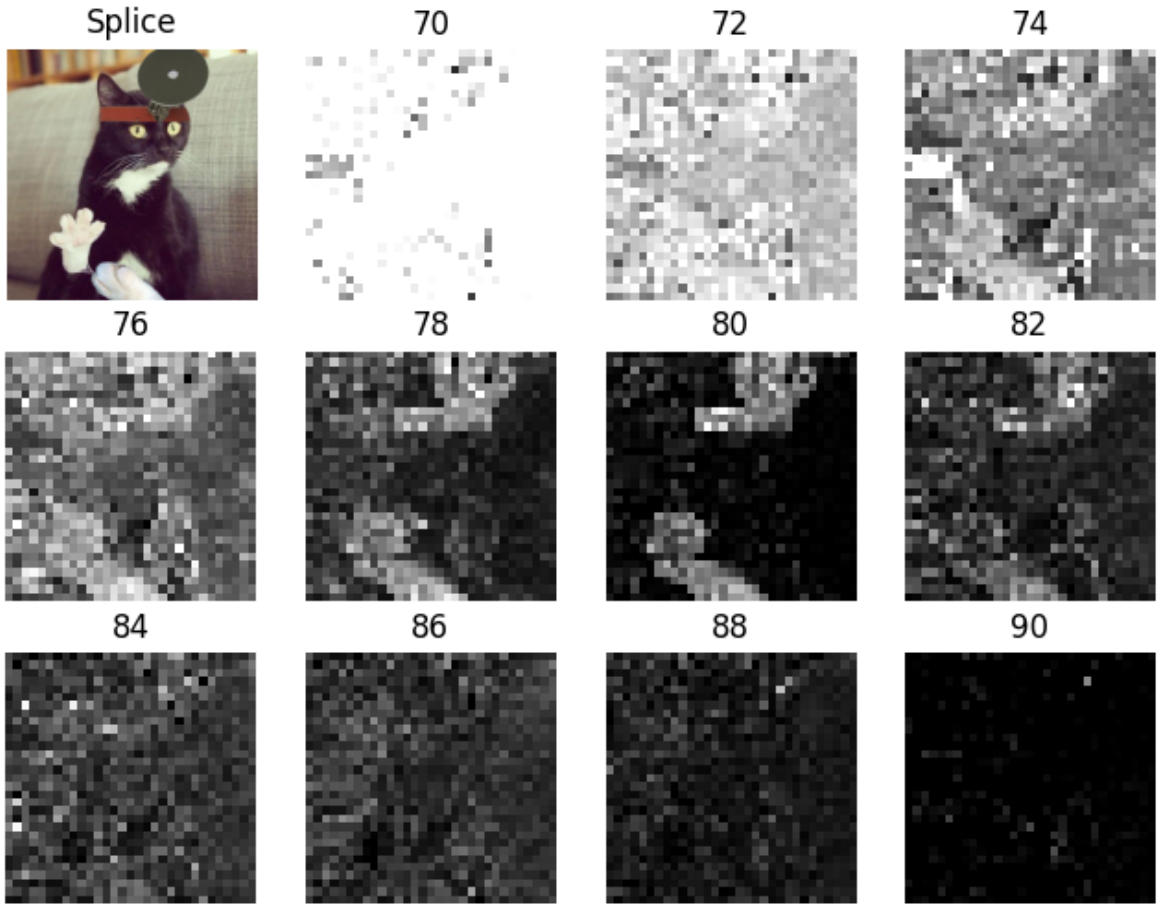

Figure 7: A modified image of a cat where various objects, such as a new paw, an operation glove around the original claw, and a doctor's head mirror, have been added. The manipulated objects are highly salient at quality 80 due to the image being saved at a higher JPEG quality than the original. The original image JPEG quality was 80%.

In figure 7, modifications have been made to a classroom picture and was saved at a higher JPEG quality than the original. The ghost map at quality 95 demonstrates that the entire classroom is of lower quality than some other regions in the image, a JPEG ghost. Specifically, the left projector screen is copied twice, the lower board is a copy of the board above it, light switches in the bottom right have been copied and an outlet to the right of the board is copied to the left. These observations strongly suggest that these objects have been manipulated, as the classroom itself is assumed to be the original part of the image. Conversely, it would make little sense if the highlighted regions in the ghost maps were authentic and the lower quality classroom was inserted into the image.

Figure 7 also showcases the user interface integrated into Sherloq. The interface includes parameters such as Lower Quality, Upper Quality, and Quality Step, which allow users to target specific JPEG qualities and construct optimal ghost plots. Additionally, Radio buttons allow the user to view the output map in grayscale or color and to include or exclude the original image in the plot. There are also offset parameters to shift the image.

A particularly useful addition is the "Next Offset" and "Previous Offset" buttons, which enable users to efficiently cycle through different offsets to search for a misaligned JPEG ghost/local minima. Plots are kept in memory to enable quick back and forth between offsets.

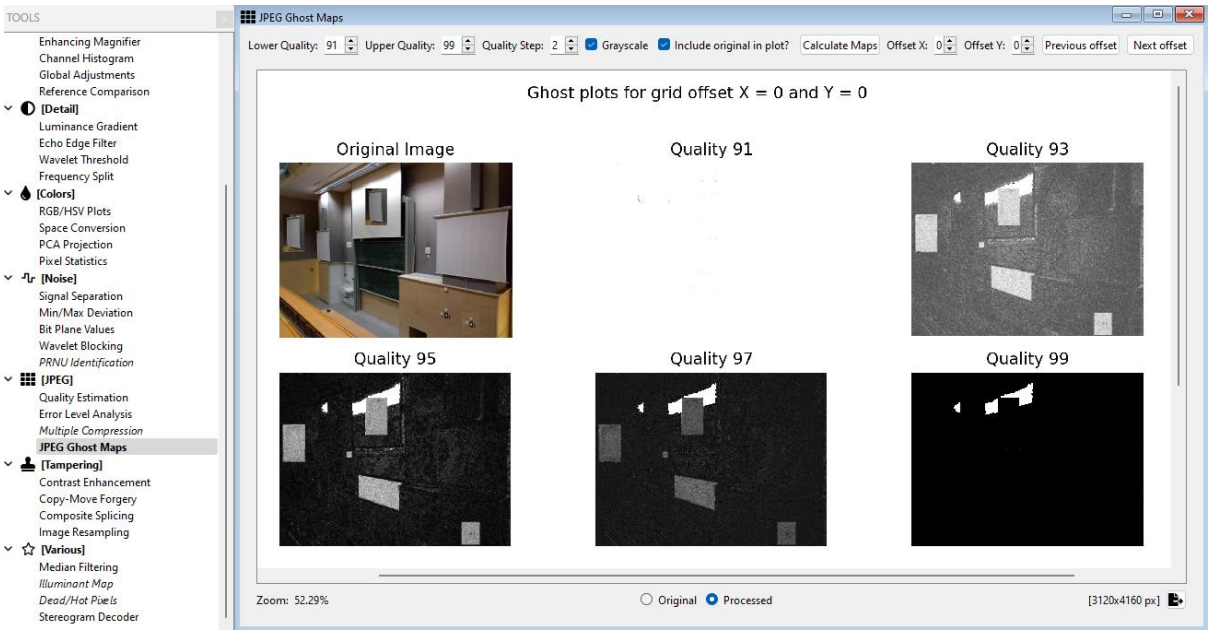

Figure 8:A classroom image with several duplicated objects, including projector screens, boards, light switches, and outlets. The manipulated regions stand out due to the image being resaved at a higher JPEG quality than the original. Notably, a persistent white region on the projector screen and top left wall represent an example of oversaturation, which must be carefully interpreted to avoid false positives when analyzing ghost maps. A quick way of differentiating JPEG ghosts from uniform regions is to check neighboring offsets. Additionally, this figure presents the user interface integrated into Sherloq.

A peculiar artifact in figure 8 is the persistent white region on the projector screen and top left of the wall. This is an example of oversaturation and can occur in authentic photographs. Therefore, oversaturated regions and uniform areas must be treated with caution when interpreting ghost maps. Uniform regions can also appear largely black because different compression levels do not significantly alter these JPEG blocks. Note that grid shifts will destroy the ghosts, but keep the uniform areas largely intact. An expected result, since the JPEG block pixel values are not changed when shifted due to the uniformity of the region.

It could be theorized that this white region was added intentionally to hide something in the image. However, a ghost map analysis cannot provide insights for this purpose. It only reveals that a region is largely uniform or oversaturated.

Key points for interpreting Ghost maps:

- This technique detects different JPEG block qualities inside an image.
- Focus on identifying highly salient regions.
- Avoid drawing conclusions from edge maps, as they typically display higher contrast and can mislead analyses. Ghosts appear and disappear near their effective quality at the correct grid offset. Position the start quality and end quality around the expected ghost.

- Less salient detections (shades of gray) are unreliable unless accompanied by a pronounced local minimum at the correct grid offset.
- Exercise caution when interpreting oversaturated pixels or uniform regions.
- Compare similar regions to reason about suspected artifacts
- Refrain from making claims if uncertain. JPEG ghosts are generally highly salient when they occur.

From a purely factual standpoint, we believe the following can be asserted about ghost maps as a means of proving image manipulation:

"This technique explicitly detects regions with different JPEG compression rates within an image. When highly salient regions are detected that are not due to uniform pixels and are part of a local minimum, their presence proves that an image was manipulated after its initial recording. These JPEG ghosts can guide practitioners in determining which regions of an image are manipulated.

Uniform regions and ghosts can be reliably differentiated by checking the JPEG grid offsets, looking for a local minimum. JPEG ghosts reveal themselves dramatically when probed at the correct alignment and quality, whereas other suspicious artifacts remain more constant across different JPEG lattices and qualities."

3.3.3 Resampling artifacts

When creating believable fakes, it is often necessary to resize or rotate parts of the image. These operations are typically performed by resampling algorithms, such as bilinear and bicubic interpolation. Detection of resampling artifacts is based on the idea that resampling algorithms alter pixel correlations in a way that makes them distinct and thus detectable compared to natural occurring correlations.

The resampling algorithm developed in this work is based on the explanations provided in the paper and a book [27,28] by Hany Farid. The code for this algorithm is available on GitHub[2], and the algorithm, along with a user interface, has been accepted as a contribution to Sherloq.

The goal of the next section is to provide a guide for understanding and effectively using the algorithm for detecting manipulations. The resampling algorithm consists of two components, which combined are used to reason about possible resizing artifacts.

1. **Probability Map:** This map reflects the probability that a given pixel's value is determined by the values of its neighboring pixels. The probability is expressed on a scale from 0 to 1, with pixel brightness indicating a high probability that the pixel value is predicted by its neighbors.
2. **Fourier Transform of the Probability Map:** High pixel correlation does not necessarily entail that the pixels are a result of interpolation. In a probability map, high pixel correlation could be a result of a uniform region, saturation or a mere coincidence, but these correlations are unlikely to have a periodic pattern if the image is authentic. A Fourier map reveals a periodic pattern by the presence of highly localized peaks.

The resampling algorithm calculates the probability map by using the Expectation-Maximization (EM) algorithm, as proposed by Dempster et al. [77]. In this approach, each pixel in an image is assumed to belong to one of two models: the first model (M1), are pixels that are correlated with

[2] *All code available at: https://github.com/UHstudent/digital_image_forensics_thesis*

their neighbors, indicating a resampling artifact; the second model (M2), are pixels that are not correlated to their neighbors.

The EM algorithm operates in two iterative steps. In the E-step, the algorithm estimates the probability that each pixel belongs to either M1 or M2 based on its correlations with neighboring pixels. In the M-step, the algorithm updates these correlations based on all samples from the E-step. This process iterates until no significant updates occur in the M-step, or is exited after a predefined number of iterations.

In practical terms, this means that an NxN mask is passed over the image, predicting correlations for each pixel based on its neighbors. In this study, all examples are demonstrated using a 3x3 mask. Pixels on the edges on an image are excluded from the probability map because they have insufficient neighbors to fit under a 3x3 mask. A 5x5 mask can also be employed, potentially detecting correlations that a 3x3 mask might miss. We integrated a 5x5 mask option into Sherloq and note that a 5x5 mask is computationally significantly more expensive. Comparing the performance of different mask sizes is beyond the scope of this study. For the purposes of demonstrating the correct implementation of the resampling detection algorithm, we will exclusively use 3x3 masks

When analyzing images using this method, the objective is to find evidence of resampling in the Fourier map. The probability map serves as input, helping practitioners to target suspicious regions and observe them in the Fourier domain. It is the presence of highly localized, periodic peaks in the Fourier map that reveals resampling artifacts.

The original paper and book conclude that a significant drawback of this technique is that it is only applicable to uncompressed TIFF images and JPEG images with minimal compression. Specifically, the book states that resampling analyses may not be effective for JPEG qualities below 90%. Therefore, the impact of JPEG compression must first be understood before this technique can be used to detect resampling artifacts.

Figure 9 demonstrates that different devices can have slightly different JPEG compression schemes that are visually distinct in a Fourier map. The general pattern of JPEG compression is both intuitive and distinct, often resembling a JPEG grid, with the cross section generally being the most pronounced. The impact of JPEG compression on the probability map is that the higher the JPEG compression rate, the more probable that a pixel is predicted by its neighbors = brighter pixels. This effect is demonstrated by comparing the probability maps in figure 10 and figure 11

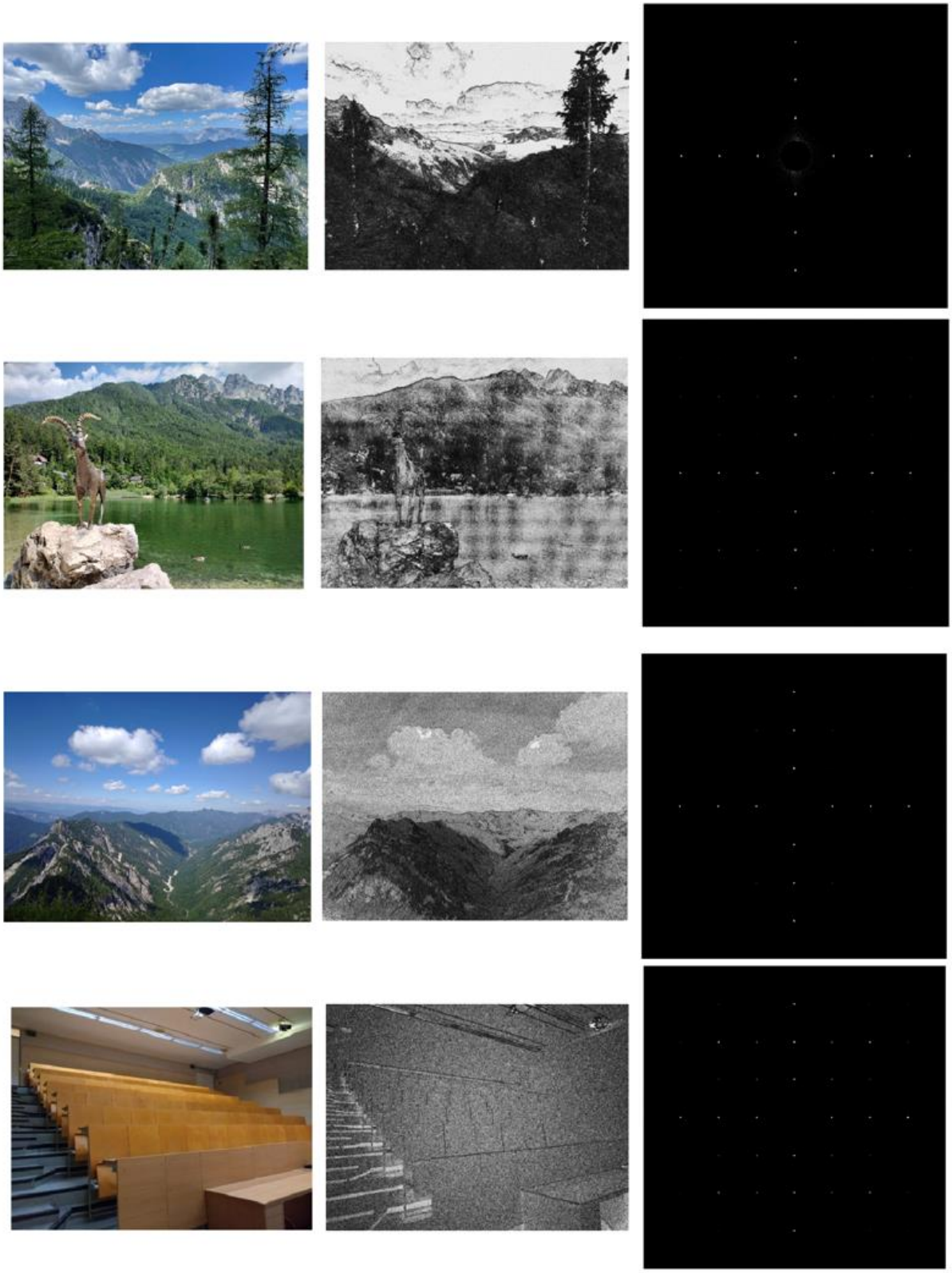

Figure 9: Four images taken by 4 different devices, top to bottom: Iphone 12, Motorola G62, Asus Camera, Moto G5, accompanied with their probability maps and Fourier map. This illustrates that different devices can exhibit slightly different JPEG compression patterns. The bottom Fourier map pertaining to the Motorola G5 camera exhibits the strongest JPEG grid structure, even though all images were recorded at quality 95%, except for the Iphone 12 image, which was recorded at quality 92%.

Figure 10 shows a manipulated image of JPEG quality 95%. A 500x500 region from the top right corner was upsampled by a factor of two using a linear interpolation algorithm and then reinserted into the bottom right corner. The probability map shows that the resized region has a higher likelihood of its pixels being predicted by their neighbors.

Other regions, such as the edges of the stairs on the left and the light near a projector in the top right, also display higher activation in the probability map. This high activation can be explained by the uniformity of the areas and their lighting conditions, leading to more predictable pixel values. Many edges of objects appear particularly dark, this occurs because edges contain rich detail of the image, leading to less predictable pixel values.

Observe the JPEG structure in the Fourier map of the global image (a). Two Fourier maps (c and d) follow the global JPEG pattern, but the Fourier map of the resized region (b) does not. What's more, four periodic dots(highlighted in red) have appeared in between the larger JPEG peaks. This is the periodic pattern that indicates this region was resampled.

In figure 11 the impact of compression is further examined. The same image was saved at a JPEG quality of 60%, and then a 500x500 section from the top right was upsampled by a factor of two using linear interpolation and reinserted into the bottom right corner.

The Fourier map (c) still reveals a periodic pattern inside the resized area that does not belong to JPEG compression. What is unexpected, is that the Fourier map (c) from the region of the camera no longer displays a grid structure similar to the overall compression scheme (a). In contrast, the seats (d) now produce a stronger grid similarity. This discrepancy could be attributed to the content's nature: the seats represent a largely uniform region with few edges, in contrast to the saturated pixels near the camera and strong edges. Ultimately, the cause of this discrepancy is not as important, what is important is that the peaks in Fourier map c do not deviate from the global structure in Fourier map a. For Fourier map b this is different, where distinct peaks are visible in between the expected JPEG compression peaks.

Figure 11 also includes the Fourier maps of the resampled region for different interpolation algorithms: INTER_NEAREST, INTER_AREA, INTER_CUBIC, INTER_LANCZOS4. form the OpenCV library. Demonstrating the impact of various interpolation algorithms on periodic peaks in the Fourier spectrum.

It is important to note that the inner peaks are also caused by interpolation in Fourier map b of figure 11. The reason for highlighting the outer peaks is that these peaks are more reliable as indicators, because peaks close to the origin could be confused with uniform or oversaturated pixels. Compare the small inner peaks From Fourier map c in figure 11 to the inner peaks caused by interpolation in Fourier map b. Although these peaks are similar, the peaks near the origin for map c result from oversaturated pixels caused by intense light.

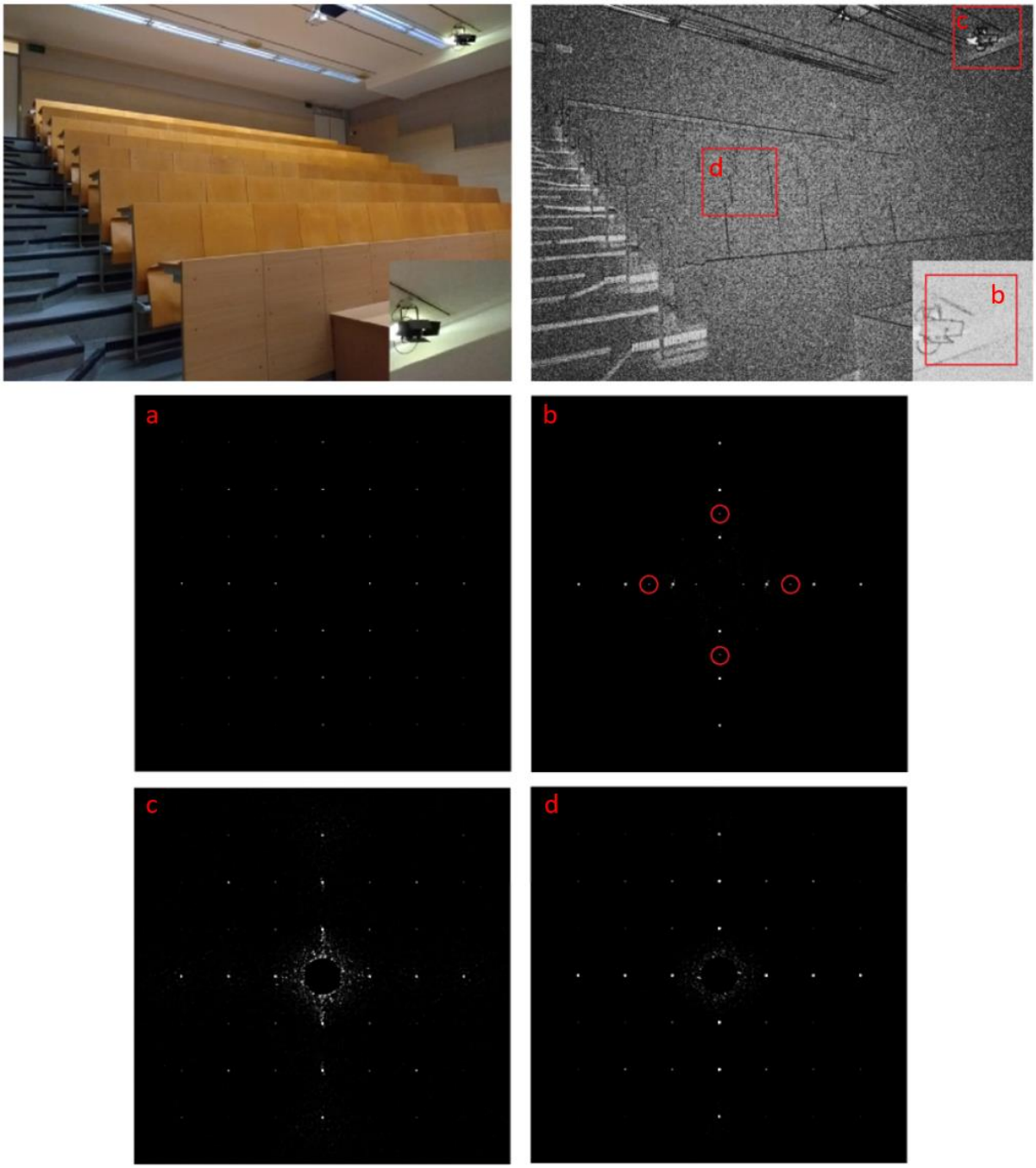

Figure 10: Analysis of a manipulated image at 95% JPEG quality where a top right portion has been upsampled by two and reinserted into the bottom right of the image. The result was then resaved at JPEG quality 100. The probability map shows a higher likelihood of predicted pixel values in the resized region. Fourier maps pertain to selected regions on the probability map (top right). a: Fourier map of global image, displaying a general JPEG structure. b: Fourier map of the resized region, displaying extra peaks circled in red and deviates from the global JPEG structure. Indicating that the region is resampled. Notice both Fourier maps c and d, which are from unmanipulated regions, follow the global JPEG pattern without extra peaks..

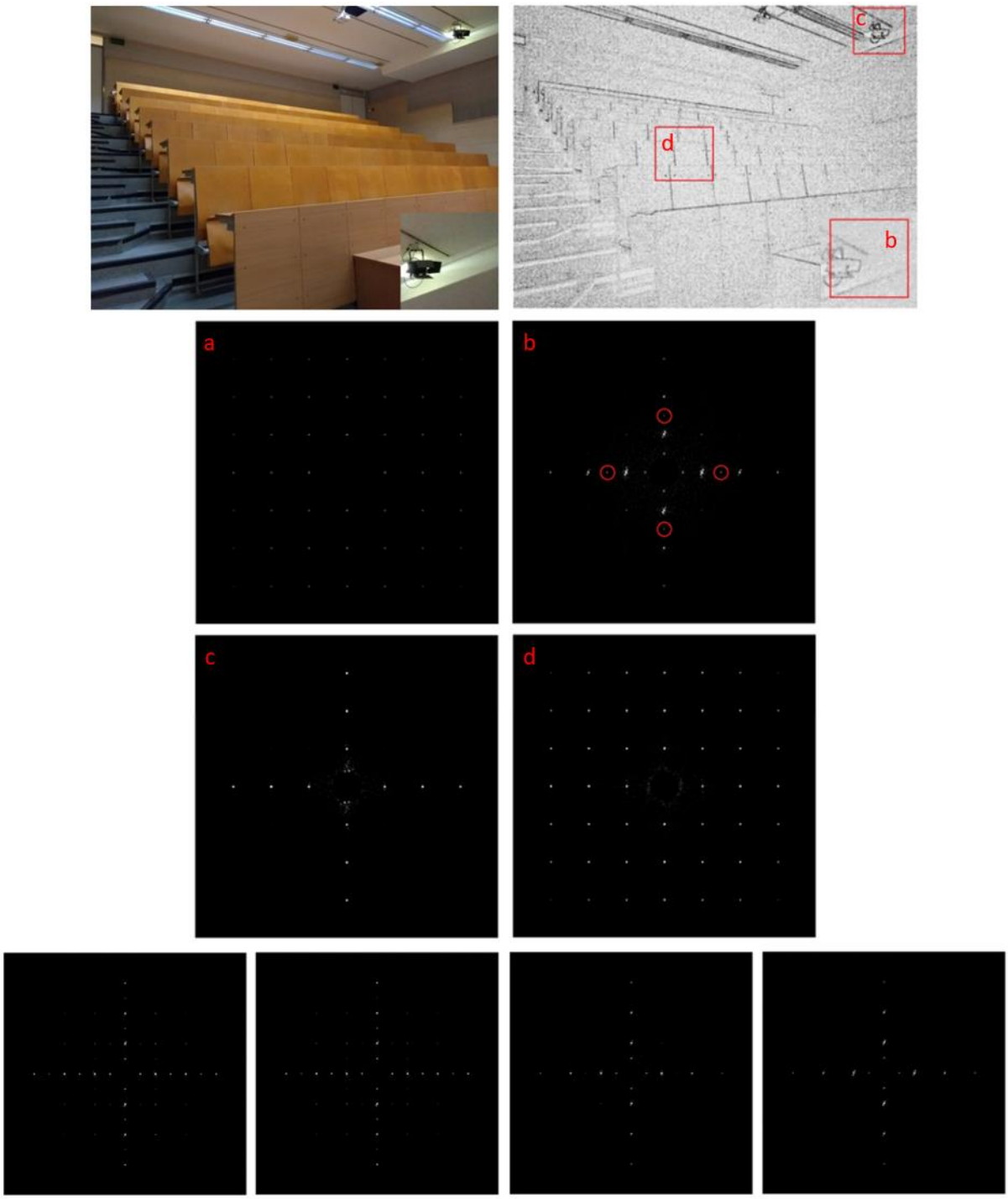

Figure 11: Analysis of a manipulated image at 60% JPEG quality where a top right portion has been upsampled by two and reinserted into the bottom right of the image. The result was then resaved at JPEG quality 100. The Fourier maps pertain to selected regions on the probability map (top right). Top left: Fourier map of global image, displaying a general JPEG structure. Top right: Fourier map of the resized region, displaying extra peaks circled in red and deviated from the global JPEG structure. Indicating that the region is resampled. Bottom left: Fourier map of the top right rectangle on the probability map. Bottom right: Fourier map of the middle rectangle on the probability map. The bottom row Fourier maps demonstrate the impact of different interpolation algorithms. Left to right: INTER_NEAREST, INTER_AREA, INTER_CUBIC, INTER_LANCZOS4 interpolation algorithms form the OpenCV library. The outer peaks are highlighted as more reliable indicators, peaks close to the Fourier map center can be more easily confused with high frequency noise or saturation in an image..

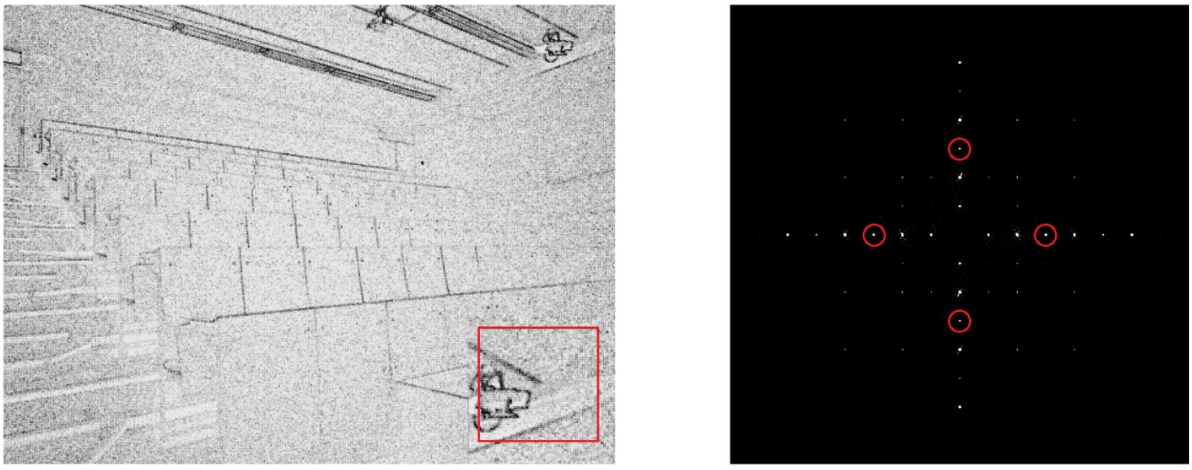

Figure 12: Analysis of a manipulated image at 60% JPEG quality where a top right portion has been upsampled by two and reinserted into the bottom right of the image. The result was then resaved at JPEG quality 60. The Fourier map demonstrates that the resampling artifact is still visible. Also the JPEG grid structure has become more visible, suggesting that linear resampling destroys parts of JPEG compression induced resampling artifacts.

In the previous examples, a part of the image was resized for different compression levels. Yet the final image was always saved at 100% quality to preserve the full impact of the resizing operation. When the image is resaved at its original quality, some unexpected results appear. Figure 12 demonstrates that the resampling artifacts are still visible, even at 60% quality. From this test, it can also be concluded that resampling (linear) possibly destroys part of the JPEG compression, as the Fourier map in figure 12 demonstrates a stronger JPEG compression grid than the same manipulated area in figure 11.

These experiments suggest that JPEG compression might not be as significant a weakness of this method as previously thought. However, this is not entirely the case. Figure 13 presents two examples where the resized region is upsampled by only 10%. When such an image is re-saved at a quality of 60%, no traces of resampling can be detected. However, when the image is saved at a 95% quality, a weak periodic signal is detectable. From these examples, it can be concluded that resampling artifacts can be identified as long as they are robust enough to withstand being overpowered by other periodic artifacts induced by operations such as JPEG compression. This observation can likely be attributed to the process of constructing Fourier maps, where it's essential to filter out noise and weaker signals in order to isolate and identify stronger periodic patterns. Without these filtering operations, an analyst cannot reliably discern periodic patterns.

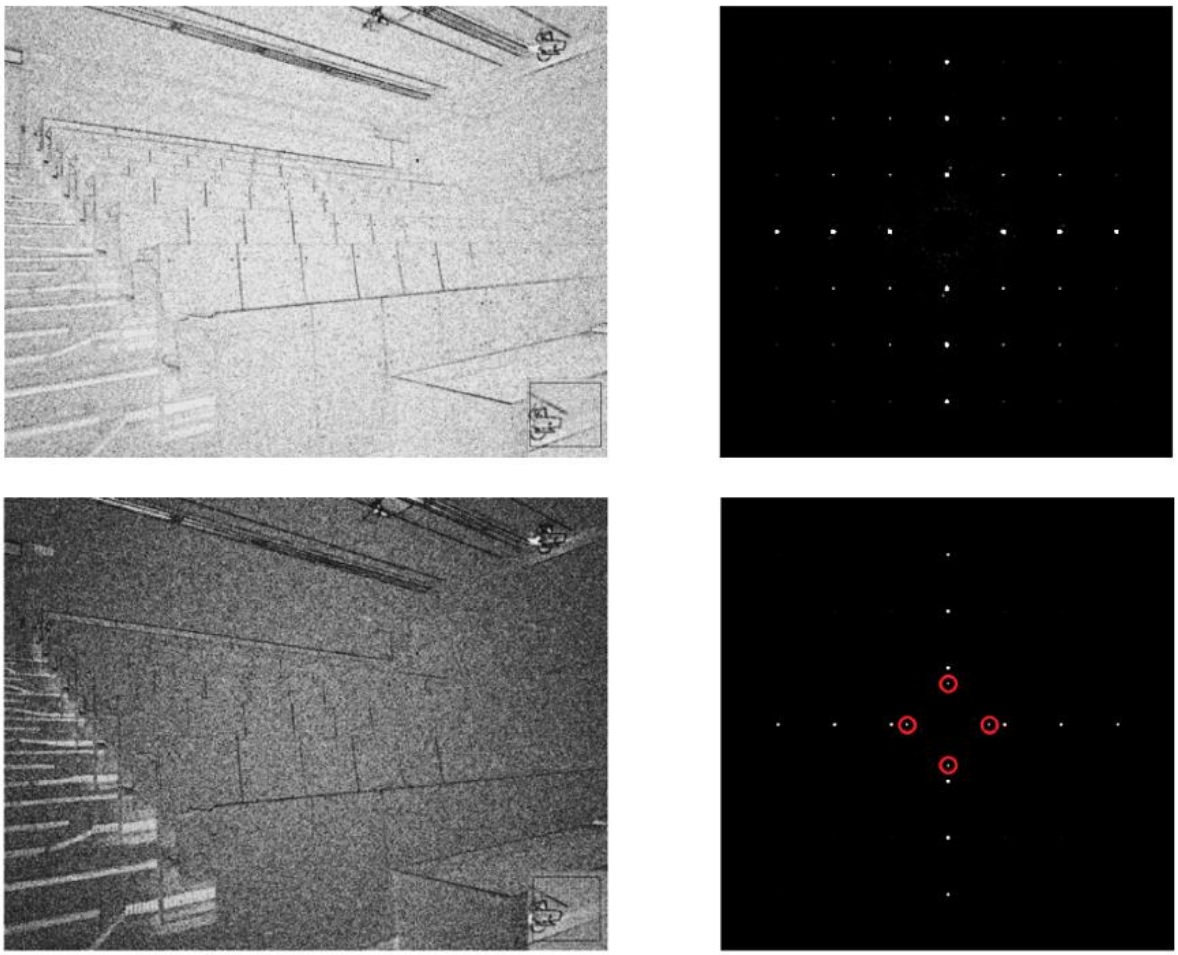

Figure 13: Two examples demonstrating the effect of subsequent JPEG compression after upsampling a region by only 10%. At JPEG quality 60% (top image), resampling traces are undetectable, but at 95% quality (bottom image), weak periodic signals emerge, indicating manipulation. This example also demonstrates the more pronounced JPEG grid structure at stronger compression levels.

3.3.3.1 Constructing Fourier maps

When processing probability maps to construct Fourier maps, several steps can be employed to visualize peaks more clearly. Both the referenced paper and book [27, 28] offer methodologies for this, and both approaches have been implemented into Sherloq. These steps can be combined for varied results and effects. While the technical aspects of each step are detailed in the respective sources, the practical implications are discussed here.

Hanning VS Rotationally invariant Window (R. I. W.)

- Hanning window: Computationally less expensive at the expense of more noise residuals.
- R. I. W.: Computationally more demanding, but generally yields sharper outputs.

Upsampling

Upsampling can help visualizing peaks because this operation shifts high frequencies, that would normally appear on the edge of the map, towards the center. This step increases computation time for subsequent steps.

Take center of the Fourier Transform:

This step takes the center of the Fourier map, potentially revealing peaks that might otherwise not have been visible. However, this can cause peaks to appear at the map's edge, possibly complicating analysis.

Highpass 1 or Highpass 2 Filter

- Highpass 1: Corresponds to the book's filter and is computationally inexpensive.
- Highpass 2: Corresponds to the paper's filter method and is computationally more demanding. Generally provides sharper filtering near the center of the Fourier map at the cost of leaving more noise near the mid frequencies.

Gamma correction

Gamma correction is an operation that scales the intensity of an image in a non linear way. This potentially highlights strong peaks, making them easier to distinguish. Yet, this step can also confound analysis by reducing weaker peaks, potentially making them invisible to the human eye.

- Paper recommendation: Gamma correction of 4, followed by rescaling the spectrum.
- Book recommendation: Gamma correction of 0,5 without rescaling the spectrum.

Rescale spectrum

Especially useful when working with high gamma corrections, because high gamma corrections tend to leave only the highest peaks visible to the human eye, which leaves little information to interpret if the spectrum is not rescaled.

Paper's recommended method:

1. R. I. W.
2. Take center Fourier transform
3. Highpass 2 filter
4. Gamma Sorrection: 4
5. Rescale Spectrum

Book's recommended method:

1. Hanning
2. Upsample
3. Highpass 1 filter
4. Gamma correction: 0,5

Our recommended method:

1. Hanning
2. Upsample
3. Highpass 1
4. Gamma correction: 4
5. Rescale spectrum

After extensive testing, the default steps in Sherloq are a combination from the book and paper steps. On average this configuration strikes a good balance between result clarity and computational demands.

3.2.3.2 Summary guide for practitioners using this technique to uncover manipulation

The general approach for utilizing this detection algorithm can be surmised in the following steps:

- **Identify the global JPEG pattern:** Analyze the entire image to understand its inherent JPEG compression structure.
- **Target Suspected Regions:** Compare Fourier maps to find a periodic pattern that deviates from the JPEG compression structure.
- **Comparative Analysis:** Compare similar regions to each other to confirm potential anomalies.
- **Local Analysis:** Conduct localized analyses of probability maps. This method is more computationally efficient and can sometimes yield better results.
- **Fourier map consistency:** Only compare Fourier maps that were constructed using the same processing steps.

Remember that periodic artifacts caused by resampling are easily disrupted by subsequent operations that modify an image's pixel intensity range (grayscale). For example, upsizing, downsizing or rotating of the entire image can destroy local resampling artifacts. Other common operations that can destroy resampling artifacts include contrast enhancement and noise addition.

The presence of resampling artifacts clearly indicates tampering, while its absence provides little evidence for the authenticity of an image.

3.2.3.3 Sherloq user interface

Figure 14 showcases the user interface developed for Sherloq. The interface includes a brief guide on how to use the tool, a selector to choose between a 3x3 Mask or a 5x5 Mask, all processing options for constructing the Fourier map with selectors, a zoom bar to scale the output and export buttons for saving the processed images.

A particularly useful feature is the inclusion of "Probability Windows" and "Fourier Windows" selectors. These options enable users to perform local analysis on specific regions of an image, which is computationally more efficient and may yield better results. The tool calculates all highlighted regions and users can quickly change these regions to target different areas of interest.

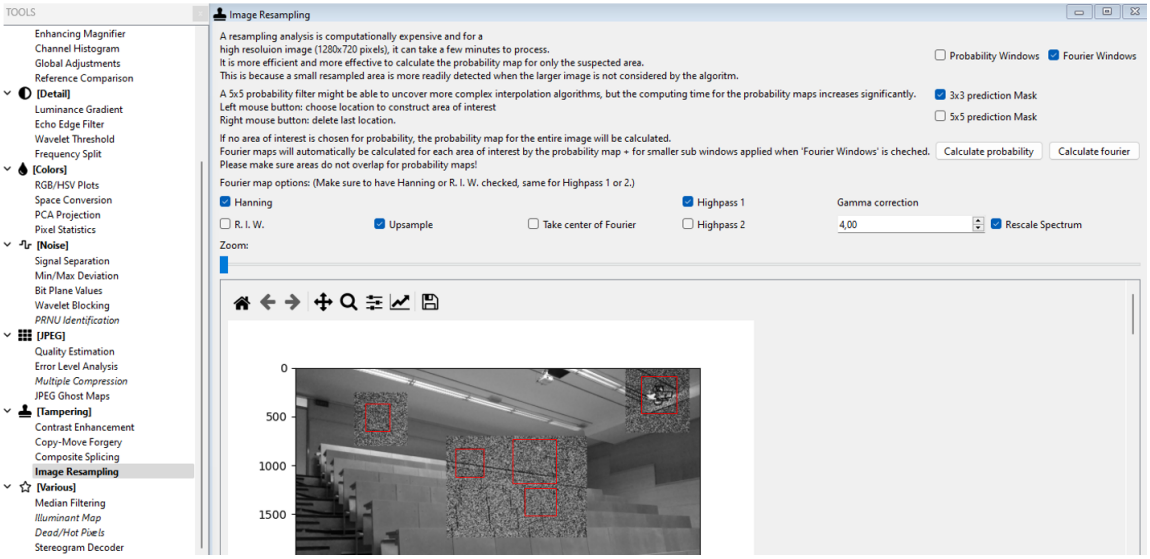

Figure 14: User interface of the resampling tool integrated into Sherloq. Demonstrating an example of local analysis using the Probability Windows and Fourier Windows feature. The red boxes have been added to highlight the selected Fourier regions, in Sherloq these are black and are sometimes difficult to observe.

3.3.4 Noise Wavelets

To avoid detection it is common that noise is added to the image, either locally or globally. This manipulation can be investigated by examining noise levels and using anomalies as evidence of tampering. The noise detection algorithm used in this work is based on the explanations provided by Mahdian et al. in their work [29].

The rest of this section will provide a brief technical insight into how this algorithm functions, followed by examples that demonstrate the functionality of the algorithm and how it can be used to detect manipulation.

The algorithm in general terms segments an image into regions based on noise levels, using wavelet decomposition and block-based noise estimation. The following steps provide a step by step guide and a brief technical explanation for the most important steps. Note that the "blocks merging" step was excluded from the final implementation. For the complete code, please refer to the GitHub [3]repository.

1. Convert the image into grayscale:
   This step isolates structural features like edges, textures, and other high-frequency components, rather than color information. It also makes the following steps more efficient.

2. Decompose the image using a 2D Discrete Wavelet Transform:
   The Daubechies 8 (db8) wavelet is chosen for this transformation, because according to literature [76] db8 is particularly effective for extracting noise characteristics, which is why it is frequently used for noise removal. In this algorithm, however, the focus is on visualizing noise, which is predominantly captured in the diagonal sub-band.

3. Wavelet blocking:
   The diagonal sub-band is divided into non-overlapping blocks. The block size can be changed: Too large reduces localization accuracy and too small affects reliable interpretability.

4. Noise level estimation:
   The noise level for each block is estimated using formula (2)
   $$\sigma = \frac{median(|HH_1|)}{0{,}6746} \quad (2)$$
   Where HH_1respresents the diagonal sub-band values per block.

The Diagonal sub-band is generally half the dimension size of the original image, further segmenting that output in blocks reduces the size of the output map significantly. It generally needs to be upscaled (or zoom in) to properly visualize the noise map. In the Sherloq implementation, the output is automatically rescaled to the original image for clarity. Figure 15 showcases the user interface developed for Sherloq.

[3] *All code available at: https://github.com/UHstudent/digital_image_forensics_thesis*

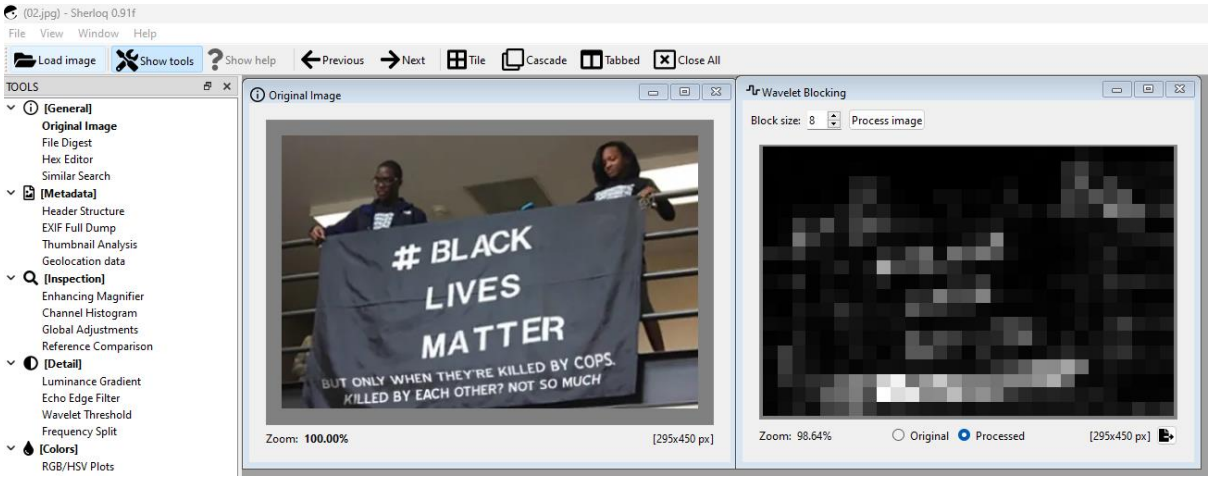

Figure 15: User interface in Sherloq for the noise wavelet algorithm (Wavelet blocking window). The block size parameter is adjustable to accommodate user preferences. The viewer interface, which includes the "Original" and "Processed" radio buttons as well as an "Export Image" button, was an existing feature in Sherloq. This interface was integrated into the noise map interface due to its practical benefits for analyzing and exporting results.

Mahdian et al. concluded that the main drawback of this technique is that authentic images can also contain different noise patterns and regions. Another limitation is that this technique is not reliable for detecting small manipulated regions, since noise is characteristic of small irregularities in an image. For these reasons, it is advised that this technique is used as a complementary technique in combination with other methods. It is difficult to assess how reliable this technique can be in the hands of an expert without performing a qualitative study. Therefore, the examples interpreting noise artifacts are merely presented as theories.

Figure 16 shows an original image of JPEG quality 95%, had its top-right region extracted, resized, and reinserted into the bottom-right corner of the image before being saved at qualities of 95% and 60%. For the JPEG quality of 95%, the manipulated region is highly salient, but this is not the case for the 60% JPEG quality. JPEG compression seems to degrade the original noise levels of the image. For the noise map of quality 60, it can be argued that the object borders inside the resampled region display weak noise levels compared to similar borders elsewhere in the image and therefore it can be labeled as an anomaly and used as proof of manipulation. How reliable this evidence might be, is not studied to our knowledge.

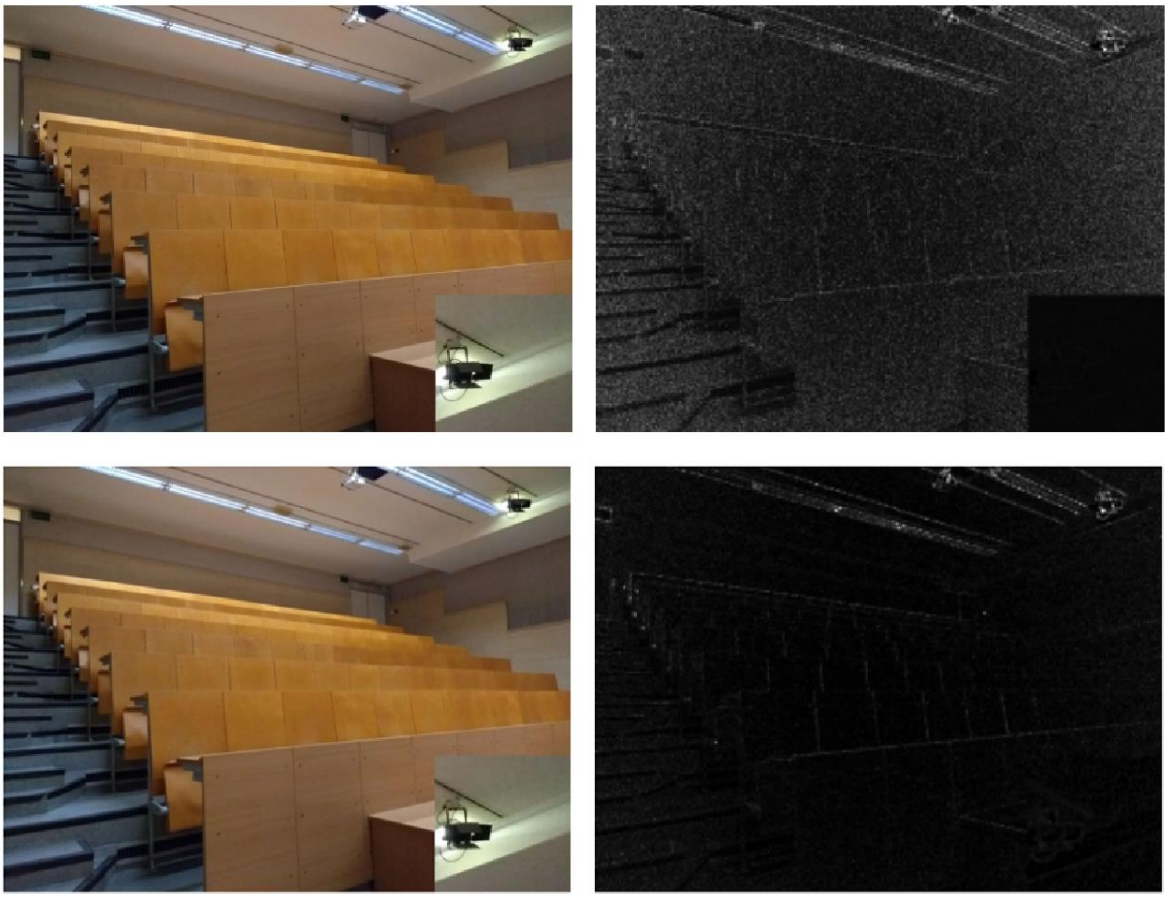

Figure 16: For an original image of JPEG quality 95% the top-right region was resized and reinserted into the bottom-right corner. The image was then saved at 95% and 60% qualities. For the 95% quality image (top), the manipulated region is a strong anomaly. For the 60% quality image (bottom), the noise levels have been degraded, making the manipulation less detectable. The weaker noise levels around object borders within the resampled region compare to similar border in the rest of the image may serve as evidence of manipulation, though the reliability of this observation has not been studied to our knowledge.

Figure 17 shows two fake images that were fact-checked by finding the original image on the web. By analyzing both original and fake images, a sense can be developed of when and how to draw conclusions with this technique. Both images will be analyzed in isolation to avoid influencing decision-making.

For the original black lives matter image[30]:

The text on the shirts of both individuals display high noise levels. This could be due to manipulation, but can also be the result of the many borders and contrast differences of the white letters on a black shirt. The regions containing the text on the flag also display higher noise levels for the same reasons. The particularly high noise levels on the shirt of the left individual is the most suspicious artifact. However, because the area is so small it is safer to ignore this. To conclude, no single region stands out as an anomaly and high noise levels can reasonably be attributed to edges or natural occurring noise. Additionally, similar regions display similar levels of noise. Without context, it is possible that all text was manipulated, or all of it is original. Noise wavelet analysis provides no strong evidence for either case. Taking context into consideration, everything appears plausible.

For the manipulated version of black lives matter:

The lower text on the flag has a higher noise level on average. The large text on the flag and the shirt imprints can be identified as similar regions on the image. Yet, their noise levels are less pronounced. Because similar regions contain both larger and smaller text, the bottom medium text imprint becomes somewhat of an anomaly. The size of the region is also non trivial. With the knowledge that resaving images always degrades image quality and can diminish or destroy original noise traces, it is reasonable to theorize that the heightened noise activation for the lower text indicates that it was added after the other text on the image - evidence of manipulation.

For the original army man image[31]:

There are no significant regions with heighted noise levels present in the image. The heightened noise levels around the soldier's shoulder are small in size and can be attributed to edges/high contrast. Nothing in this image is particularly suspicious. The absence of any noise in the background can be attributed to poor image quality.

For the manipulated version of the army man:

Much like the original image there are no large regions of suspicion. The upper shoulder still shows similar activation and can be dismissed for the same reasons as in the original. The stitch containing the text "Doing the work of" also shows heightened noise levels. The region is small and there are borders/high contrast in this region. Similar regions in the image are the watermarks "trinixy.ru" in the bottom right corner and to the left above the shoulder. These regions also show some activation and that is to be expected. The idea that these watermarks were inserted into the image after its original recording is trivial, but that doesn't necessarily entail that the patch under the American flag is also manipulated. It is tempting to label the patch as manipulated. After all, it makes little sense for an American soldier to be wearing this. However, the noise wavelet analysis can not provide definitive evidence.

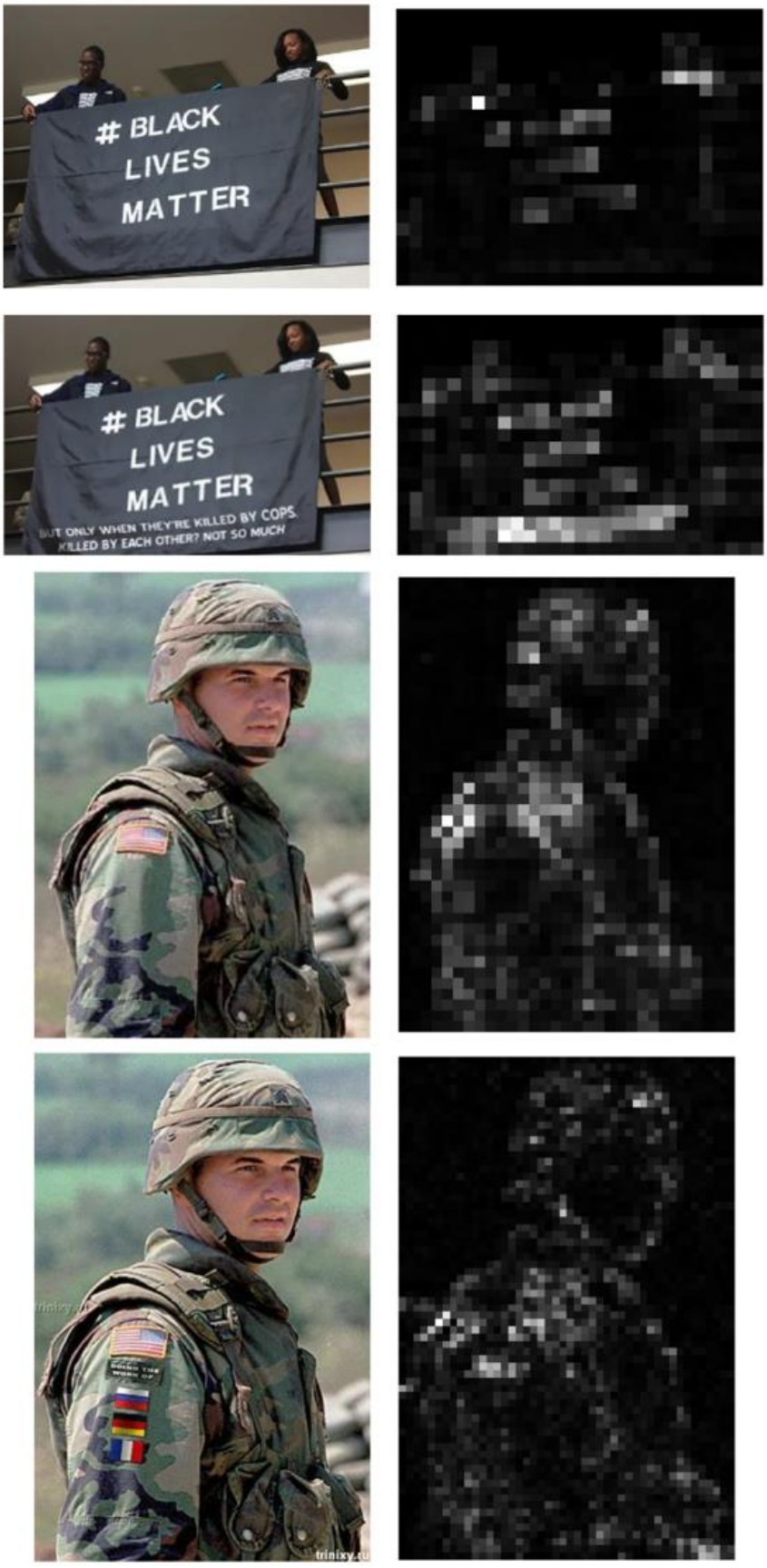

Figure 17: The first and third row show an original image with their corresponding noise wavelet maps, followed by a manipulated version on the row below.

3.4 Quantitative study between AI and traditional techniques

In the remainder of this work we will compare the three traditional localization algorithms – JPEG ghost maps, Probability maps and Noise wavelet maps against a state-of-the-art AI network; MM-Fusion. While probability maps should be used in combination with Fourier maps to identify resampling artifacts, we have included them in this study as a standalone technique. To our knowledge, Probability maps have not been quantitatively analyzed in this context, making it particularly interesting to explore their performance and assess their viability for detecting manipulations on their own. Another important detail for this study is that the ghost map algorithm will not examine offsets for misaligned ghosts. This decision is due to the significant increase in computational time required to check all offsets, which would result in a 64-fold increase in processing time.

We start by explaining our method followed by the results and a discussion. Then we will suggest future research and surmise the most important findings of this work.

3.4.1 Evaluation method

Our methodology is inspired by Zampoglou et al. [32]. Their work evaluated traditional localization algorithms on a large scale. Datasets with ground truth masks marking the manipulated area were collected in order to evaluate the algorithms. The main reason for this approach is that we're interested in localization. Without masks it would be impossible to reliably determine if a localization is correct.

For evaluation, the values under the ground truth mask are compared to the values outside the mask. Depending on the threshold, the image will then be classified as authentic or manipulated. Two statistical methods are used for this comparison: the absolute median difference [33] and the two-sample Kolmogorov-Smirnov statistic (K-S statistic) [25,32].

The absolute median difference is calculated by finding the median value of the pixels inside and outside the mask, then their absolute difference is taken.

The K-S statistic is a more sophisticated evaluation metric and assesses whether two samples have come from the same distribution. It is defined as:

$$k = max_u|C_1(u) - C_2(u)| \ (1)$$

Where $C_1(u)$ and $C_2(u)$ are the cumulative probability distribution function of the two samples.

A k value of 1 indicates the samples are completely distinct. While a value of 0 indicates the values come from the same normal distribution. In practice the K-S static was calculated using the 'ks_2samp' function [34] from the SciPy library.

In order to estimate false positives it is necessary to evaluate original, unmanipulated images. Yet there exists no ground truth mask in such cases because there is no manipulation to detect. This work will follow in the footsteps of [32] by creating an 'random' test mask where the center region of an image will be compared against the rest of the image.

A more advanced approach that would reduce the bias introduced by using the ground truth mask in the validation step would be to develop an algorithm that analyzes the output map of a detection method and determines if a region is manipulated. However, this approach requires knowing the specific threshold for each algorithm to identify manipulated regions. Furthermore, such an algorithm would require fine tuning for each detection method. For instance, characteristics of manipulated area's in JPEG ghost maps differ from those in noise wavelet maps and probability

maps. Creating and testing these algorithms for reliability is a substantial task and thus left for future work. For these reasons, we adhere to established methodologies in our comparative study.

3.4.2 Datasets

In selecting datasets, we prioritized those that were available to us and could be processed within a reasonable time frame. All selected datasets include ground truth masks highlighting the manipulated area. Table below surmises the datasets used in this study.

Table 1: Benchmark image datasets and their characteristics. Every datasetcontains ground truth mask that highlights the manipulation.

Dataset	Number of Images		Format
	Real	Fake	
Columbia [39]	183	180	.tif
IMD2020 [37]	414	2010	.JPG (JPEG);PNG
In the Wild [39]	0	201	.JPG (JPEG
CocoGlide[41]	512	512	.PNG (with possible JPEG history
IFS training set [42]	1050	450	.PNG (with possible JPEG history)
Coverage [36]	100	100	.tif

IMD2020

The IMD2020 dataset [37] includes a subset called “Real-life Manipulated Images” consisting of 2010 manipulated images and 414 real images sourced from the internet. More specifically, the images were sourced from the PS-Battles Dataset [38] and custom masks were created for the manipulated images. The PS-Battles dataset is a collection from the subreddit r/photoshopbattles where Photoshop enthusiasts manipulate images based on community requests. In this context, “Real-life” manipulated images refers to “uncontrolled” manipulation and “real” means “the original requested image”.

This dataset models how real-world manipulated images might be encountered. It features manipulated images from anonymous individuals from all over the world using different software and exhibiting varying degrees of skill, resulting in a high variety of manipulations. Figure 18 shows a few example images in this dataset.

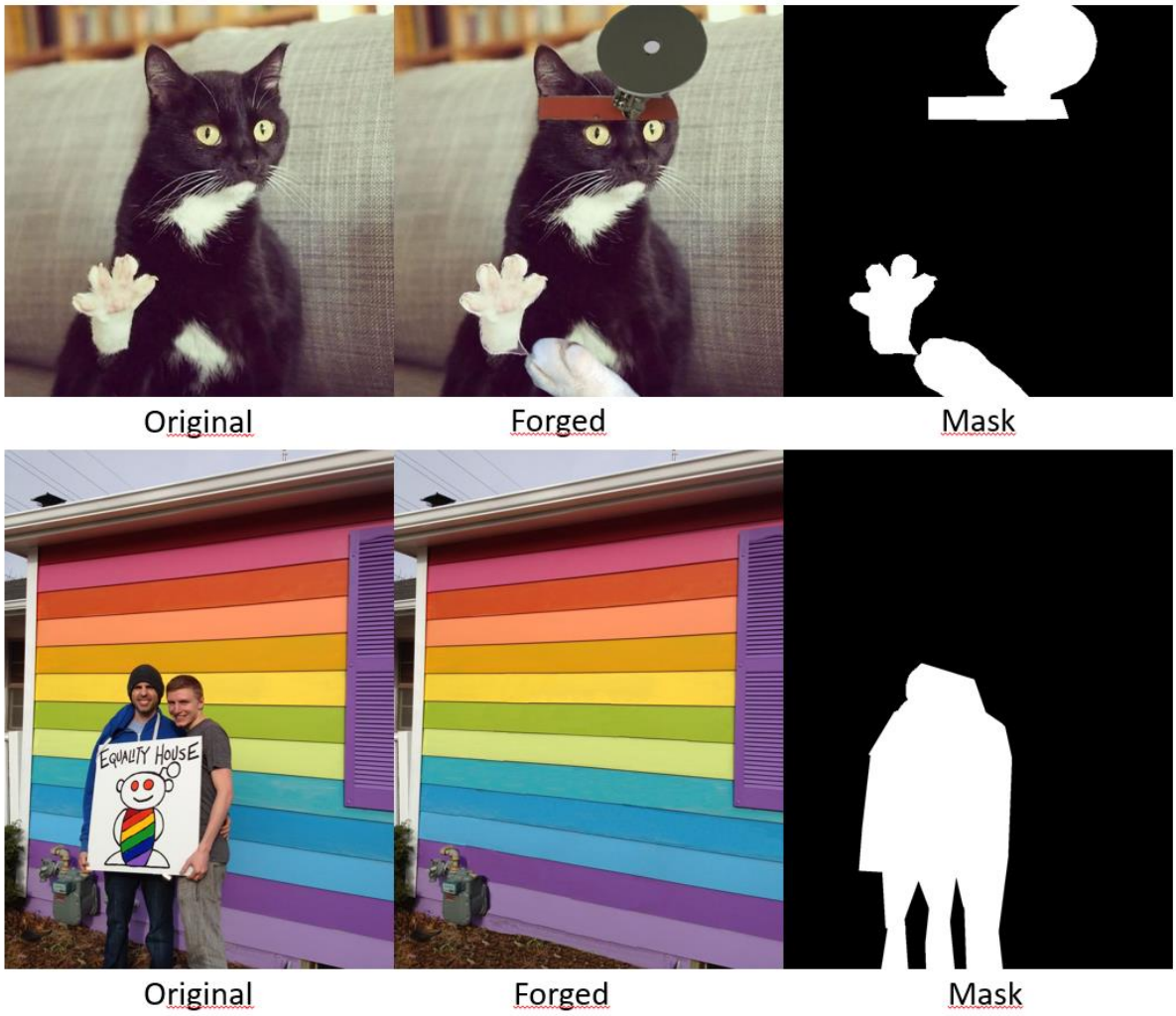

Figure 18: Sample images from the IMD2020 dataset. Showing an original, forged and mask imge that highlights the manipulated area.

In the Wild

The “In the Wild” dataset is similar to IMD2020. The authors aimed to collect a dataset that diversified the number of origins for the manipulated images in an attempt to represent how such images are encountered on the internet. The dataset consists of 201 images collected from “THE ONION” website [43] and the subreddit r/photoshopbattles. The authors manually crafted approximate ground truth masks using the source image whenever it was available. Figure 19 shows a few examples from this dataset.

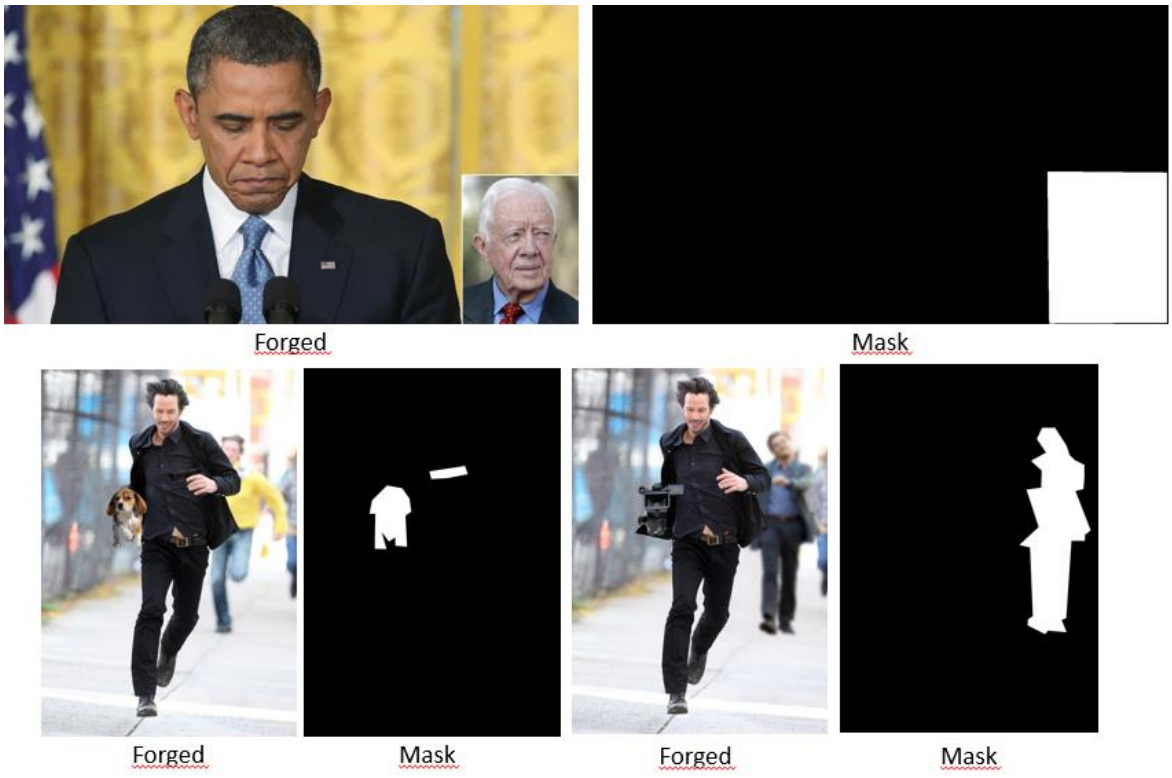

Figure 19: Sample images from the In the Wild dataset. Showing a forged and accompanying mask image highlighting the manipulated area.

Columbia

The Columbia 'Uncompressed Image Splicing Detection Evaluation Dataset' consists of 183 authentic and 180 manipulated images. The authentic images are uncompressed and captured using several different camera models. The manipulated images are spliced versions of the authentic images and no post processing has been done. Different types of ground truth masks are provided to locate the splices. Figure 20 shows some examples of this dataset.

In this study, precise evaluation masks were created by converting the original red/green masks in the dataset to black/white ground truth masks. Figure 20 illustrates the results of this conversion operation, denoted as Mask'.

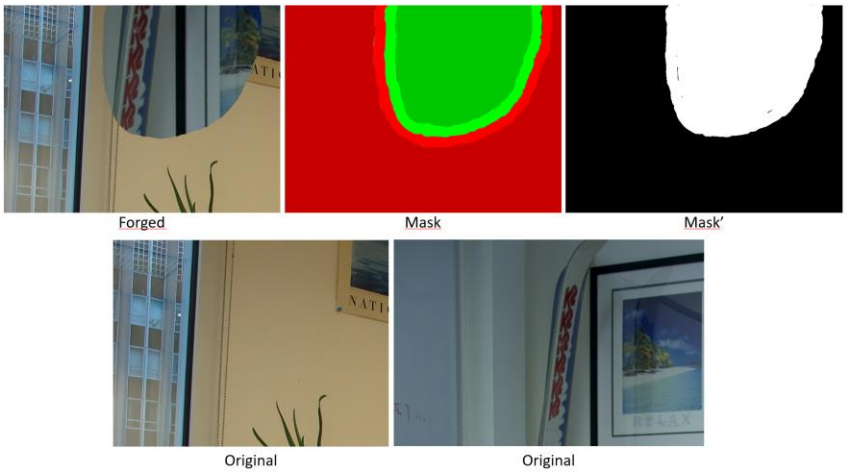

Figure 20: Sample images from the Columbia dataset. Two original images and a splice of these images are shown. The masks accompanying the forged images in the dataset were converted to black and white masks for evaluation purposes. The result of this conversion is Mask'.

CocoGlide

The CocoGlide dataset [41] contains 512 real and 512 manipulated images. The real images are 256x256 pixel crops from images in the COCO validation set [44]. For each real image, an object and a text prompt were created and these were fed to GLIDE [45]. The resulting 512 manipulated images contain synthetic objects of the same type as those in the original images. Some examples are shown in figure 21.

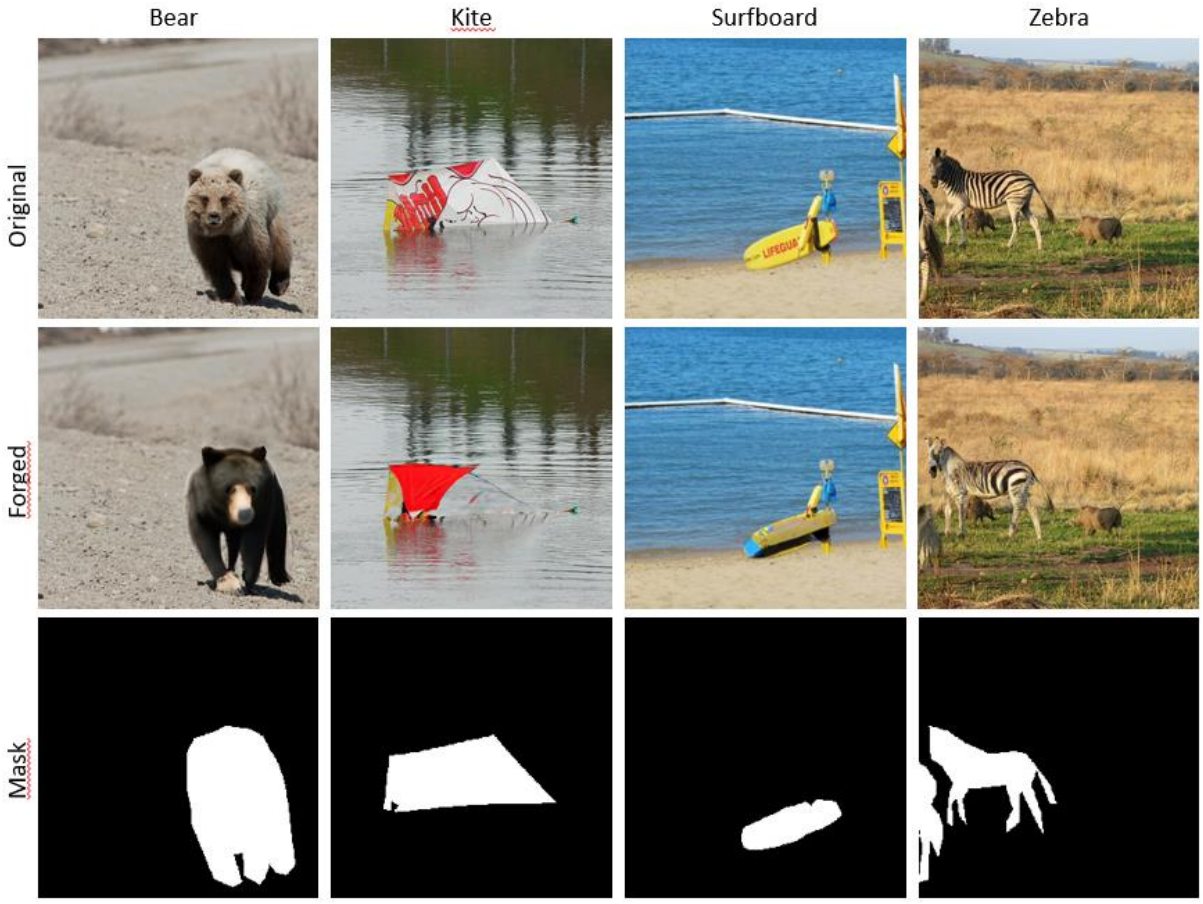

Figure 21: Sample images from the CocoGlide dataset. Showing an original, forged and mask pair images that highlights the manipulated area. Each column contains the prompt text that was used for GLIDE to inpaint the mask region

IFS

The IEEE information Forensics and Security Technical Committee (IFS-TC) organized the first international competition for digital image forensics in 2013 [42]. The organizers provided a training set of 1050 original images and 450 manipulated images accompanied with a ground truth mask. The manipulations in this dataset are realistic and created by forgers or various skill levels. Unlike the other datasets, the masks in this dataset highlight the manipulated area in black instead of white. This difference does not affect result calculations. Figure 22 shows some examples from this dataset.

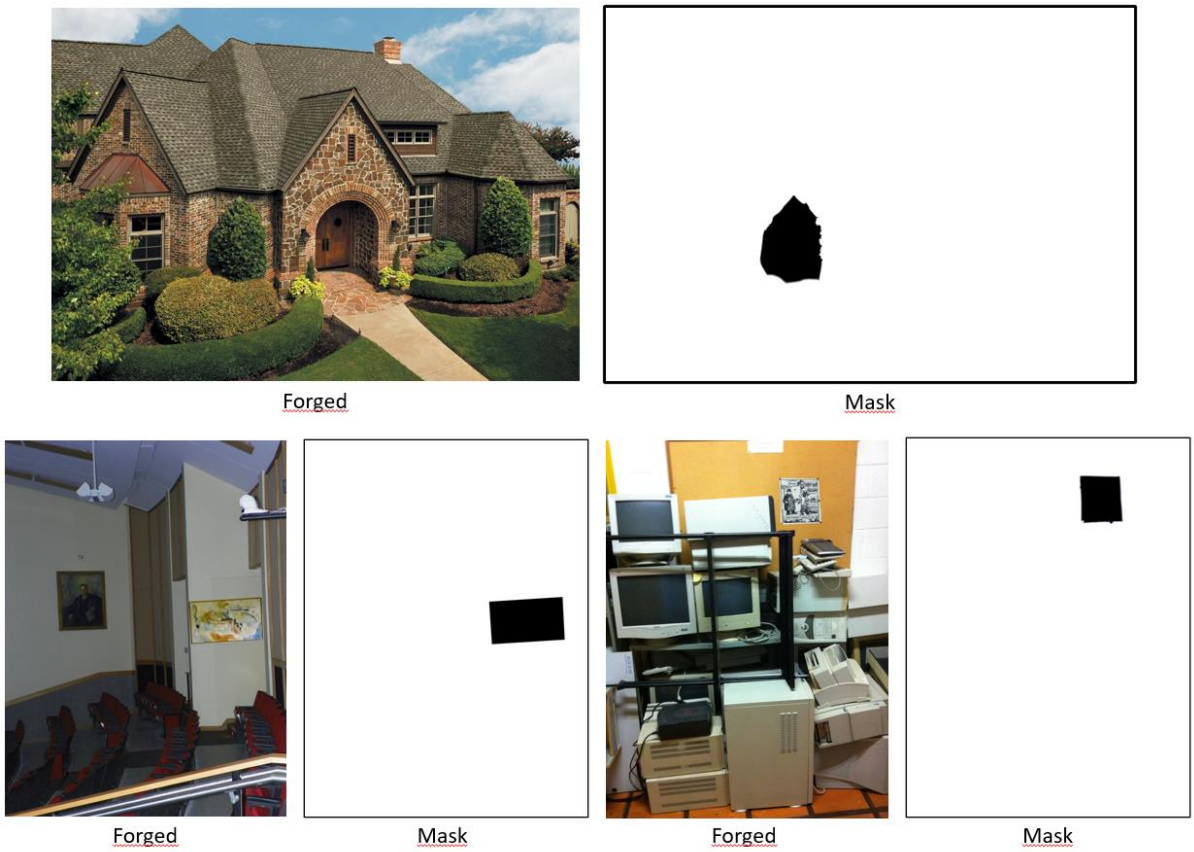

Figure 22: Sample images from the IFS training dataset. Showing forged and accompanying mask that highlights the manipulated area. The border surrounding the mask images was added for clarity and is not part of the original dataset.

Coverage

The Coverage dataset [36] is a copy-move forgery database comprising 100 original and 100 manipulated images. All images are uncompressed .tif files and contain similar but genuine objects. Two masks accompany each manipulated image: one mask locates the original object that was copied and the second highlights the location where the copy was inserted into the image. The primary manipulation in this dataset is the copy-paste of an object, sometimes slightly resized and/or rotated. Some examples are shown in figure 23.

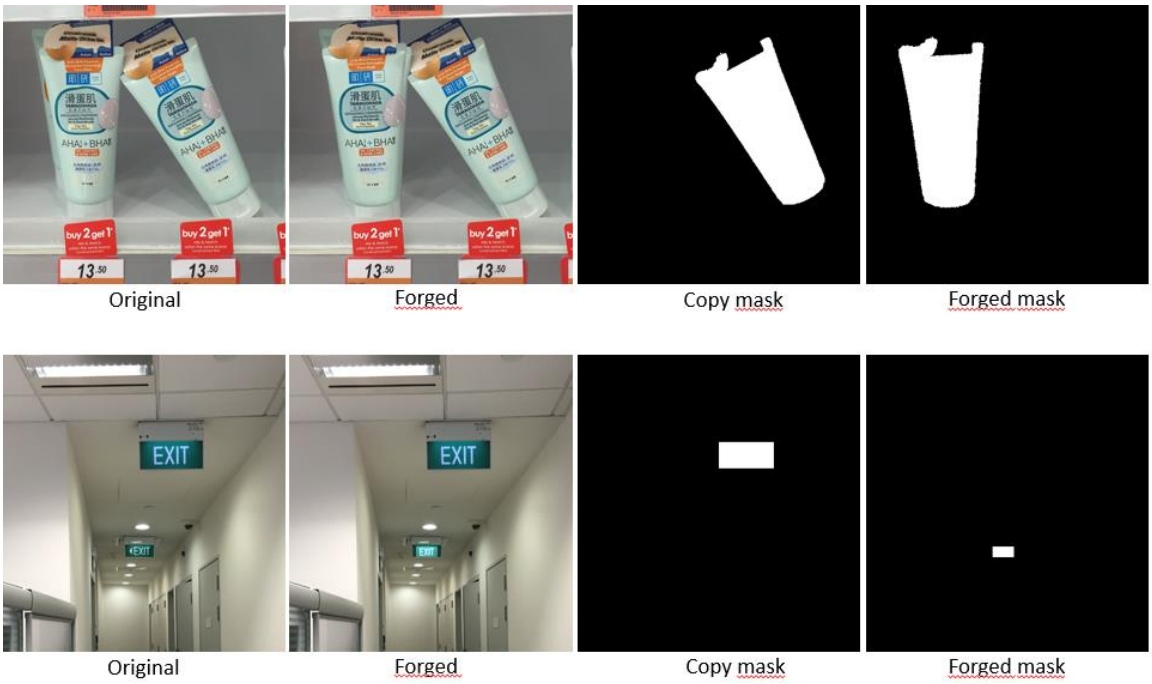

Figure 23: Sample images from the Coverage dataset. Showing an original, forged and two accompanying mask images. The Copy Mask highlights the original object that was copied and the Forged Mask highlight the region where the copy was inserted.

3.4.3 Quantitative Results

This section will discuss the results of the quantitative study, two comparison approaches will be presented, first the Receiver Operating Characteristic (ROC) curve is calculated for each algorithm per dataset. Second, overlap statistics will be calculated between the traditional algorithms and MM-Fusion.

Following this, we will qualitatively analyze the data, highlighting the most notable findings and discuss the significance of the results.

3.4.3.1 Receiver Operating Characteristic Curves

Figures 24,25,26,27,28 and 29 show the performance for each algorithm on datasets IMD2020, In the Wild, Coverage, Columbia, CocoGlide and IFS respectively. For each algorithm the ROC curve is shown for the K-S statistic and absolute median difference classifiers. The AUC has also been calculated for each ROC curve and thresholds for False Positive Rates (FPR) 0, 5, 10 and 20% are highlighted. FPR thresholds above 20% become generally uninformative.

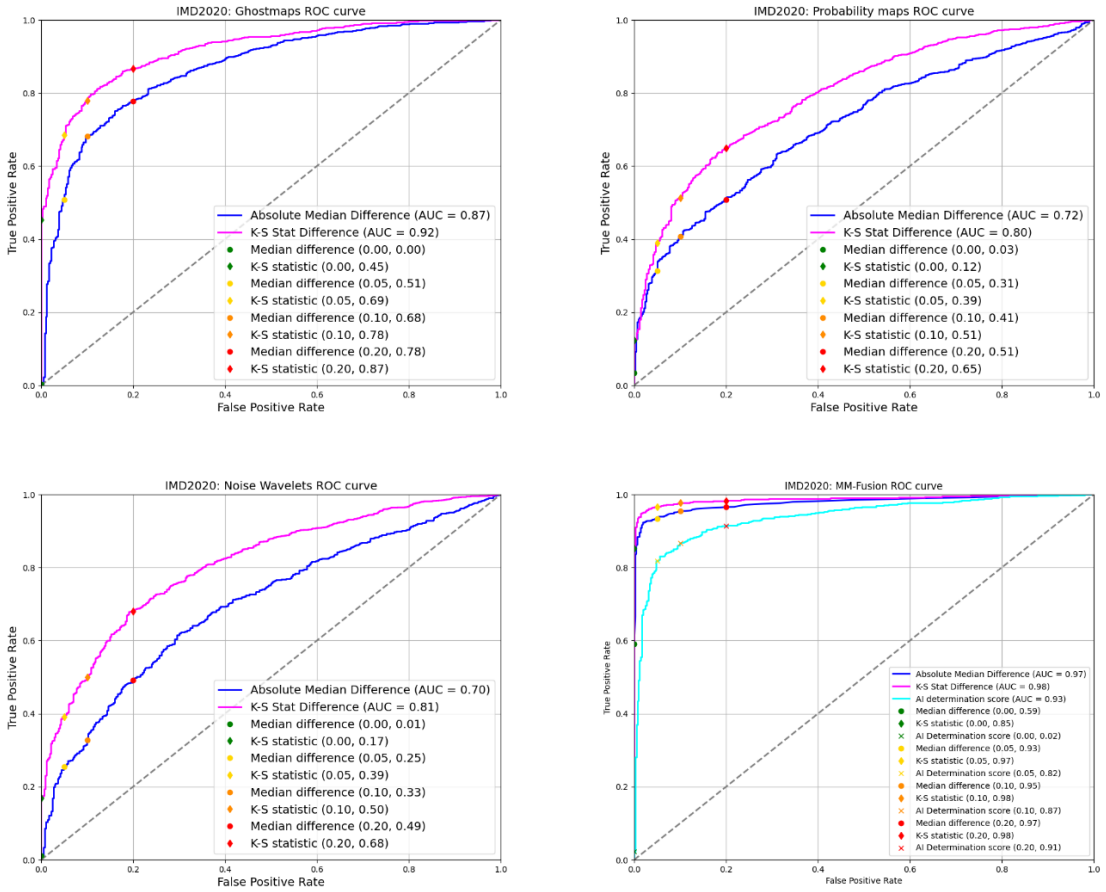

Figure 24: ROC curves for each algorithms' performance on the IMD2020 dataset shown separately for the K-S statistic and absolute median difference classifier.

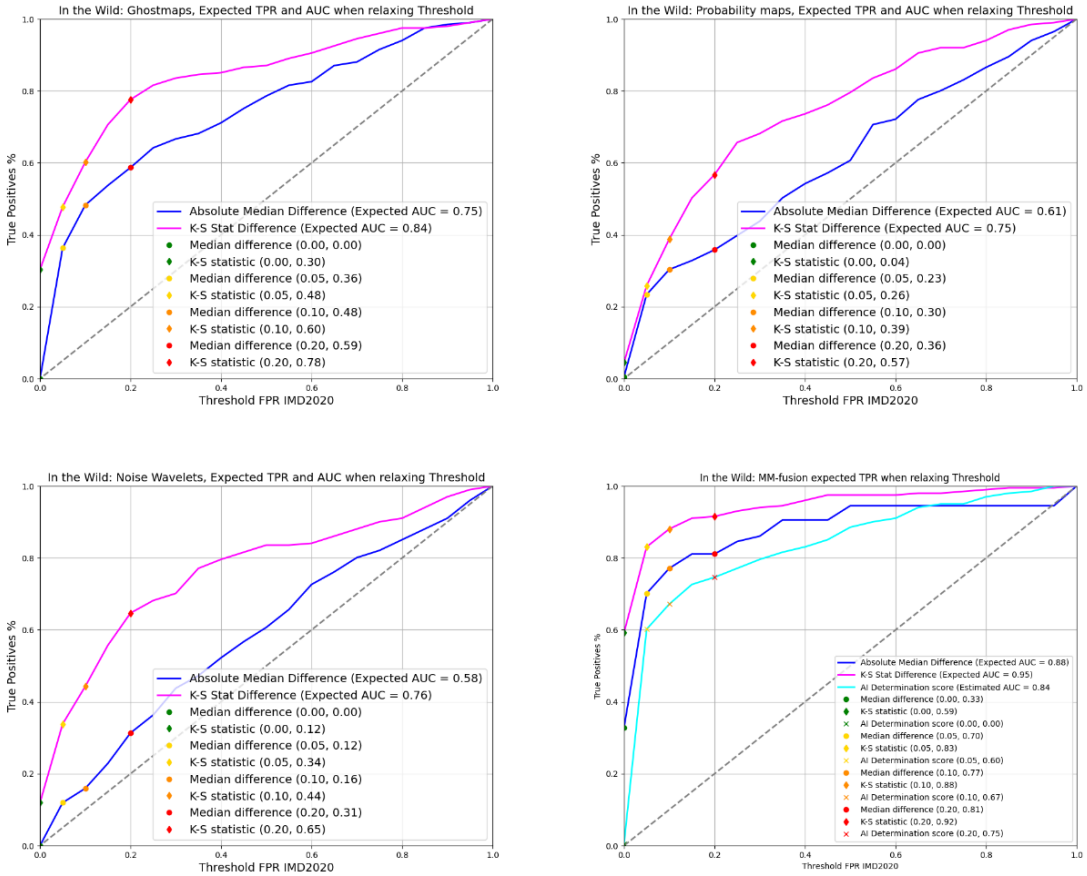

Figure 25: ROC curves for each algorithms' performance on the In the Wild dataset shown separately for the K-S statistic and absolute median difference classifier. Because the In the Wild dataset does not contain original images, the expected FPR thresholds from the IMD2020 dataset are applied and the AUC has been calculated using the trapezoidal rule.

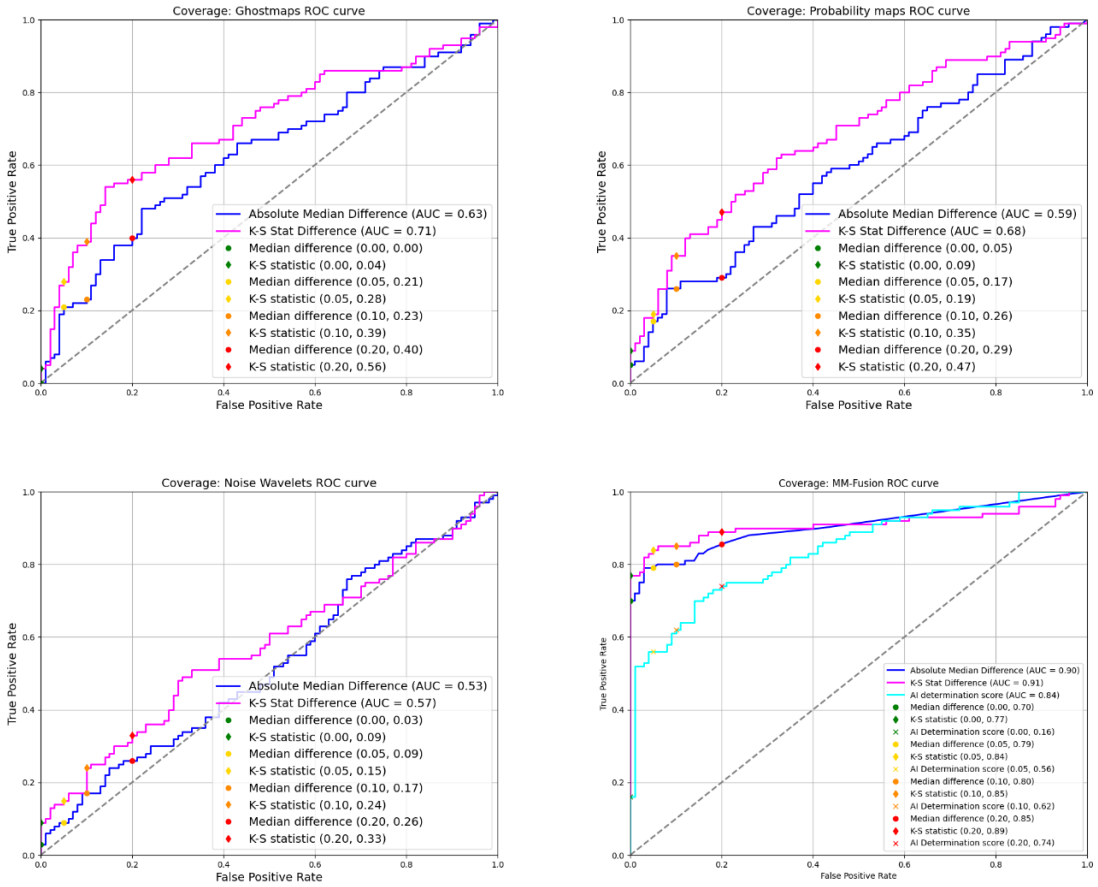

Figure 26: ROC curves for each algorithms' performance on the Coverage dataset shown separately for the K-S statistic and absolute median difference classifier.

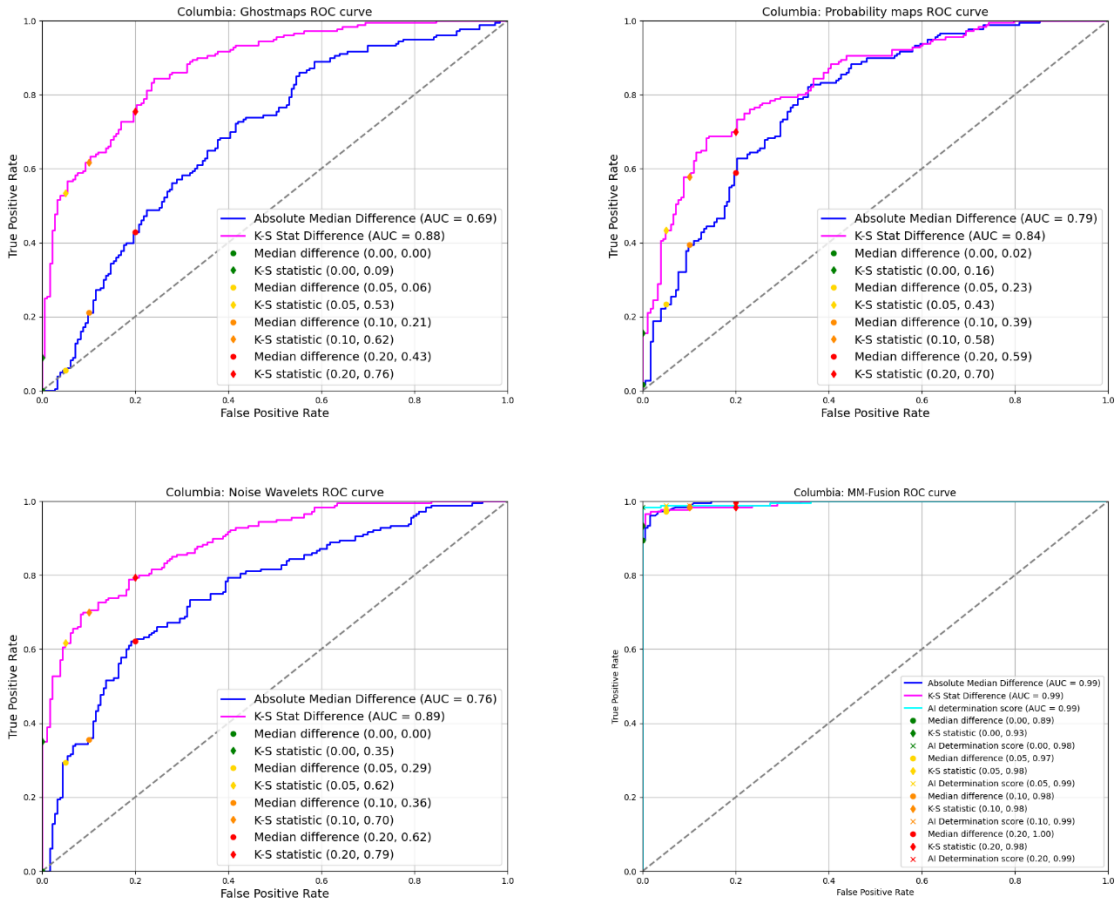

Figure 27: ROC curves for each algorithms' performance on the Columbia dataset shown separately for the K-S statistic and absolute median difference classifier.

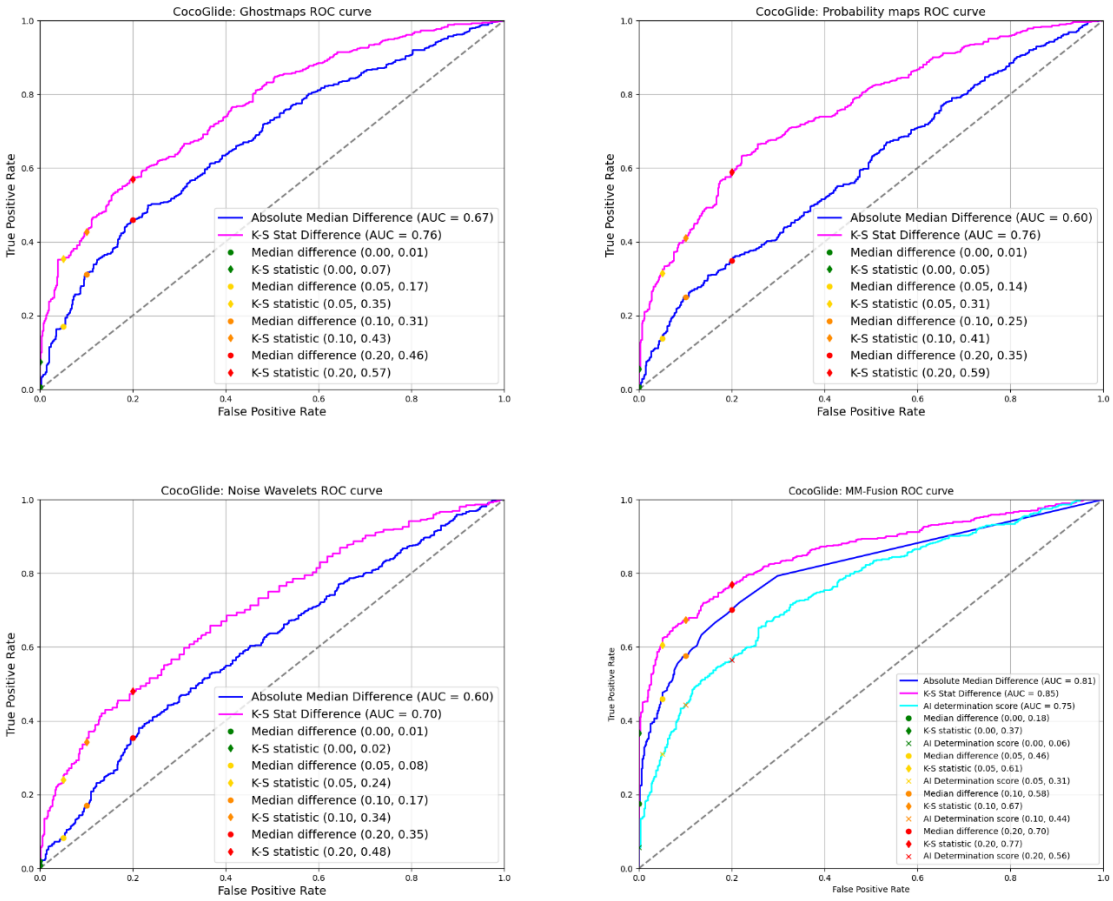

Figure 28: ROC curves for each algorithms' performance on the CocoGlide dataset shown separately for the K-S statistic and absolute median difference classifier.

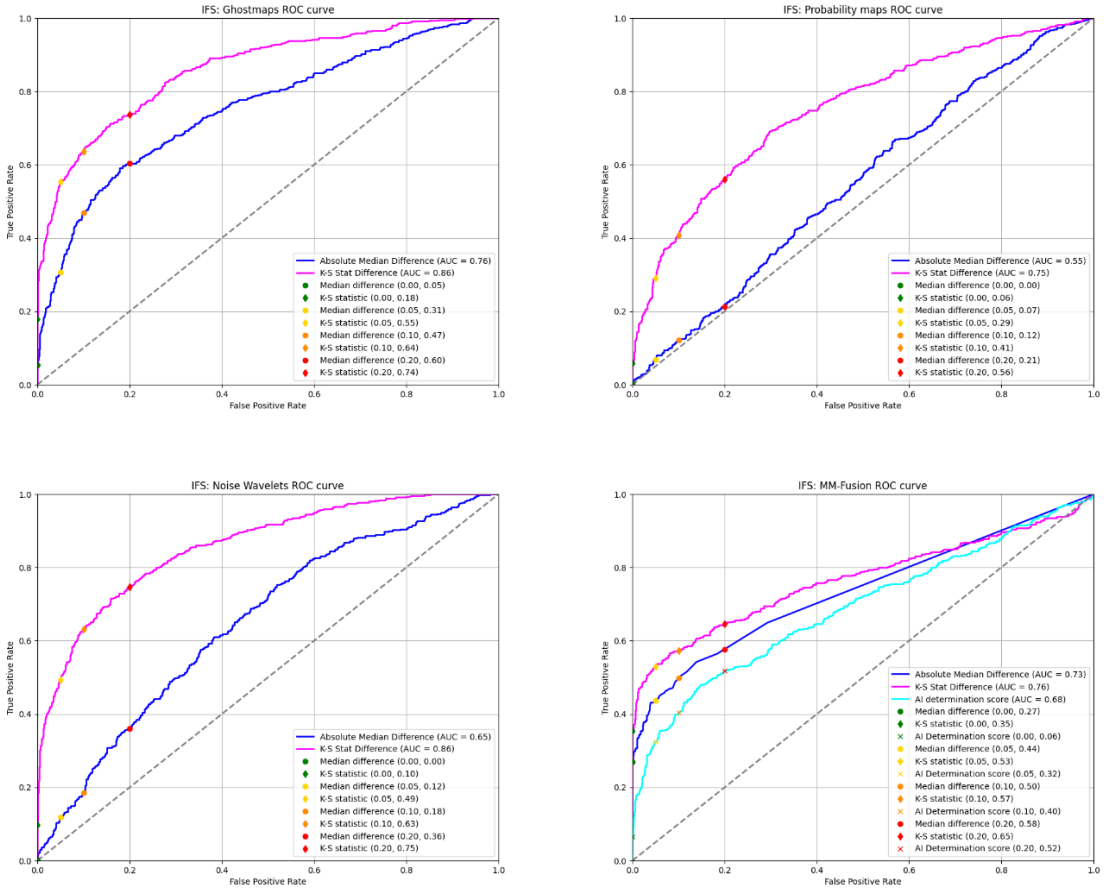

Figure 29: ROC curves for each algorithms' performance on the IFS training dataset shown separately for the K-S statistic and absolute median difference classifier.

3.4.3.2 Absolute median difference vs K-S statistic

The results show that the AUC for the K-S statistic is on average 9,42% higher than the absolute median difference's AUC. There is considerable overlap between the two predictors at every FPR level. At FPR = 0% the K-S statistic generally overlaps the absolute median predictor entirely or almost entirely. Specifically, between 60% and 100% of absolute median difference detections are also detected by the K-S statistic for each method. Specific overlap results at key FPR levels can be found in the overlap result files on GitHub[4]. They have not been extensively included in a table here because they do not provide extra useful insights.

A theoretical and practical example best demonstrates the difference between these two predictors.

Consider the following collections of pixel values inside and outside the mask:

Values_inside_mask = [0, 0 ,0 ,255 ,255] -> median = 0

Values_outside_mask = [0, 0 ,255 ,255 ,255] -> median = 255

Absolute median difference = abs(0 - 255) = 255

K-S statistic = 0,2

[4] *All code available at: https://github.com/UHstudent/digital_image_forensics_thesis*

This theoretical example demonstrates how the absolute median can change drastically based on just one pixel. In contrast, the K-S statistic indicates a low likelihood that both collections are dissimilar. Since most masks are not pixel perfect translations of the manipulations they highly, it follows that the volatile nature of the absolute median predictor is less reliable for these test conditions.

For a practical example shown in the figure 30, the following statistics were calculated for the probability map:

- Median inside mask = 0,0
- Median outside mask = 230,925
- Absolute median difference = 230,925
- K-S statistic = 0,415

When inspecting the probability map in combination with the mask, it becomes apparent that the high absolute median difference threshold is a result of specific pixel distribution rather than an accurate representation of how dissimilar the area under the mask is compared to the rest of the image. The K-S statistic indicates the two samples are more similar than they are distinct, which aligns more closely with visual observation.

For the noise map the following statistics were calculated:

- Median inside mask = 61,928
- Median outside mask = 34,274
- Absolute median difference = 27,654
- K-S statistic = 0,435

Here the absolute median value indicates the regions inside and outside the mask are similar, which is in line with the visual observation. The K-S statistic indicates the same observation.It is important to note that the quantitative approach lacks critical reasoning. When comparing the region under the mask with similar objects, regions and lighting conditions, both for the probability and noise map, the manipulation begins to stand out as an anomaly and warrants further investigation.

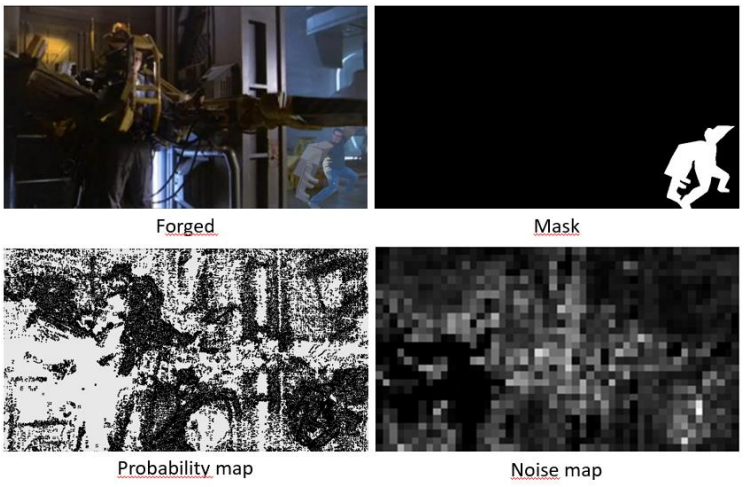

Figure 30: Practical example for comparing the K-S statistic and the absolute median difference as a classifier. For the probability map, the absolute median difference (230.925) is influenced by specific pixel distribution, suggesting that the dissimilarity between the masked and unmasked region is high. In contrast, the K-S statistic (0.415) suggests greater similarity, which aligns more closely with visual inspection. For the noise map, both the absolute median difference (27.654) and the K-S statistic (0.435) indicate that the regions inside and outside the mask are similar, corroborating visual observations.

In conclusion, the K-S statistic is more suitable for determining if two samples are similar compared to the absolute median difference classifier. This is because the inclusion or exclusion of just a few pixels has the potential to significantly change the absolute median difference classifier.

3.4.3.3 Comparison of detection accuracy with True Positives and False Positives metrics

The MM-Fusion model was trained on the IMD2020 dataset [23], hence it is observed that MM-Fusion performed exceptionally well on this dataset, achieving an AUC of 0,93 for the AI determination score. What is interesting is that the ghost maps K-S statistic classifier achieved an AUC of 0,92. Suggesting that this method is nearly as effective as the AI's detection score classifier.

In order to compare MM-Fusion against a collection of traditional techniques, the following metrics were calculated for each dataset at FPR = 0% and 5%, the results are shown in table 2.

1. Total unique detections for traditional techniques:
    - Combined for the K-S statistic, from which a TPR was derived.
    - The absolute median difference was excluded due to its inferior performance.
2. AI model TPR:
    - The TPR from the AI determination Score was used because in a blind testing scenario this would be the used metric to evaluate an output.
3. Total Unique False Positives for Traditional Techniques:
    - Combined for the K-S statistic, from which a FPR was derived.
    - The combined FPR is higher than the stated FPR because the traditional techniques do not share the same false positives.

At first glance, these statistics suggest that for 0% false positives, the combined traditional techniques outperform a state-of-the-art AI network in terms of detecting and localizing manipulated regions in images by an average of 17% more true positives for the each dataset. For the IMD2020, In the Wild and the IFS datasets this difference increases to 10 times more true positives for 0% false positives. This finding could be used to critique the validity of AI networks. After all, AI networks lack explainability compared to traditional techniques. If three techniques combined do a better and more reliable job (explainability) of detecting manipulated images at 0% FPR, it raises the question of how much better four, five or more techniques might perform.

Unfortunately, these statistics should be interpreted with the utmost caution. As we analyze and qualitatively reflect on this study, it becomes clear that these statistics are not useful for drawing meaningful conclusions. They are included here to emphasize the need for critical evaluation of perceived performance as indicated by statistics.

Table 2: Comparison of MM-Fusion and traditional techniques at FPR = 0% and 5%. Metrics include total unique detections for traditional techniques combined for the K-S statistic, and AI model TPR using the AI determination score. The table also shows total unique false positives for traditional techniques combined for the K-S statistic. The combined FPR is higher than the stated FPR due to non-shared false positives among the different algorithms. These statistics suggest traditional techniques outperform the AI model in terms of true positives at 0% FPR, particularly for the IMD2020, In the Wild, and IFS datasets, where the difference can be up to 10 times more true positives. More detailed data for other FPR levels can be found on the GitHub page in the result files.

		Columbia		IMD2020		In the Wild		CocoGlide		IFS		Coverage		AVG	
	FPR	0%	5%	0%	5%	0%	5%	0%	5%	0%	5%	0%	5%	0%	5%
MM-Fusion	TP	**0,983**	**0,989**	0,023	**0,817**	0,005	**0,602**	0,057	0,311	0,064	0,324	0,120	**0,560**	0,209	0,601
	FP	0,000	0,038	0,000	0,048	/	/	0,000	0,049	0,000	0,050	0,000	0,060	0,000	0,049
Traditional Techniques	TP (Unique)	0,394	0,767	**0,535**	0,795	**0,348**	0,592	**0,117**	**0,549**	**0,693**	**0,724**	**0,160**	0,380	**0,375**	**0,635**
	FP (Unique)	0,000	0,104	0,000	0,099	/	/	0,000	0,094	0,000	0,110	0,000	0,100	0,000	0,101

3.4.3.4 Shortcomings of validation methods & quantitative analysis limitations

The ROC curves of the test results suggest that most methods have merit as a reliable classifier on the Coverage and Columbia datasets. An exception is the noise wavelets technique for the Coverage dataset, which underperforms. This suggests that detecting copy move forgeries with the noise wavelets technique is unviable.

Let's surmise the performance (K-S statistic) of the classical techniques for these datasets.

Coverage K-S statistic AUC's:

- Ghostmaps = 0,71
- Probability maps = 0,68
- Noise wavelets = 0,57

Columbia K-S statistic AUC's:

- Ghostmaps = 0,88
- Probability maps = 0,84
- Noise wavelets = 0,89

These results are unexpected and contrary to our predictions. How is possible that ghost maps, a technique designed for JPEG compression analysis, appears to be reliable for detecting manipulations in uncompressed datasets? The answer can be found in the method used to account for false positives. In our approach a "random" mask was applied to original images. Even though others have used this method in the past, they failed to properly highlight its shortcomings.

Figure 31 demonstrates how mask placement significantly misguides the results of this testing setup. For an image pair (original + copy move) in the Coverage dataset, a K-S stat of 0,260 was calculated for the ghost map output of the original image using a 'random' square mask. For the forged image a K-S statistic of 0,772 was calculated using the "forged" mask. Based on these testing data, the results for the coverage dataset were calculated. When inspecting the ghost map outputs for both images, there was no visual difference to the naked eye. This prompted us to test a more realistic approach. Instead of using a random square mask, we applied both the copy mask and the forged mask to the original image ghost map output. The resulting K-S statistics were 0,490 and 0,704 respectively. A significant difference.

We decided to re-run the calculations using the Copy mask for evaluating the original images, because the copy mask always encapsulates an object in the original image. This choice was also appropriate for this specific dataset, as this compares the original object in an original image with the forged object in the forged image. The results for this new evaluation strategy are called "Coverage2" and can be seen in figure 32.

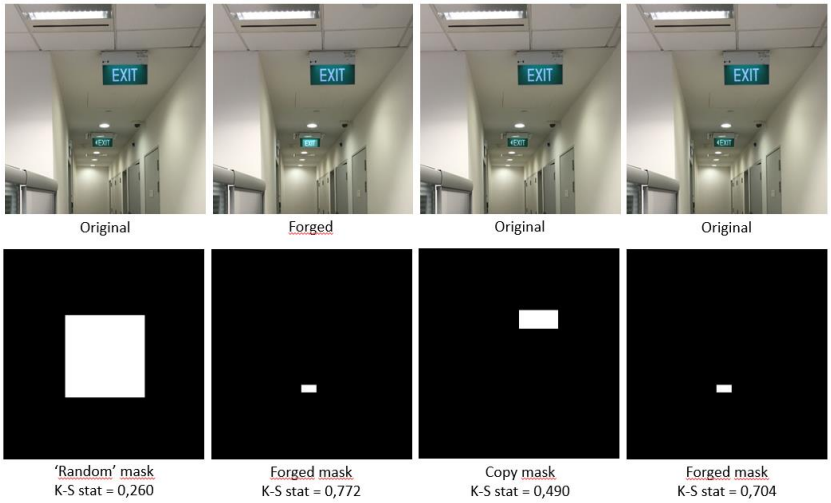

Figure 31: Demonstration of how mask placement misguides False Positives in the test results. For an image pair in the Coverage dataset, a K-S statistic of 0.260 was calculated for the ghost map output of the original image using a 'random' square mask, while the forged image ghost map output had a K-S statistic of 0.772 using the "forged" mask. This difference in K-S statistic despite there being no visual differences in the Ghostmap outputs, suggest that map placement is of critical importance. This suggestion is further supported by the K-S statistic differences when the copy mask and forged mask were applied to the original image ghost map output, resulting in K-S statistics of 0.490 and 0.704, respectively. The difference in K-S statistics of 0,772 and 0,704 for the same area in two different ghost map output despite there being no visual difference is due to the fact that these differences are imperceptible to the human eye.

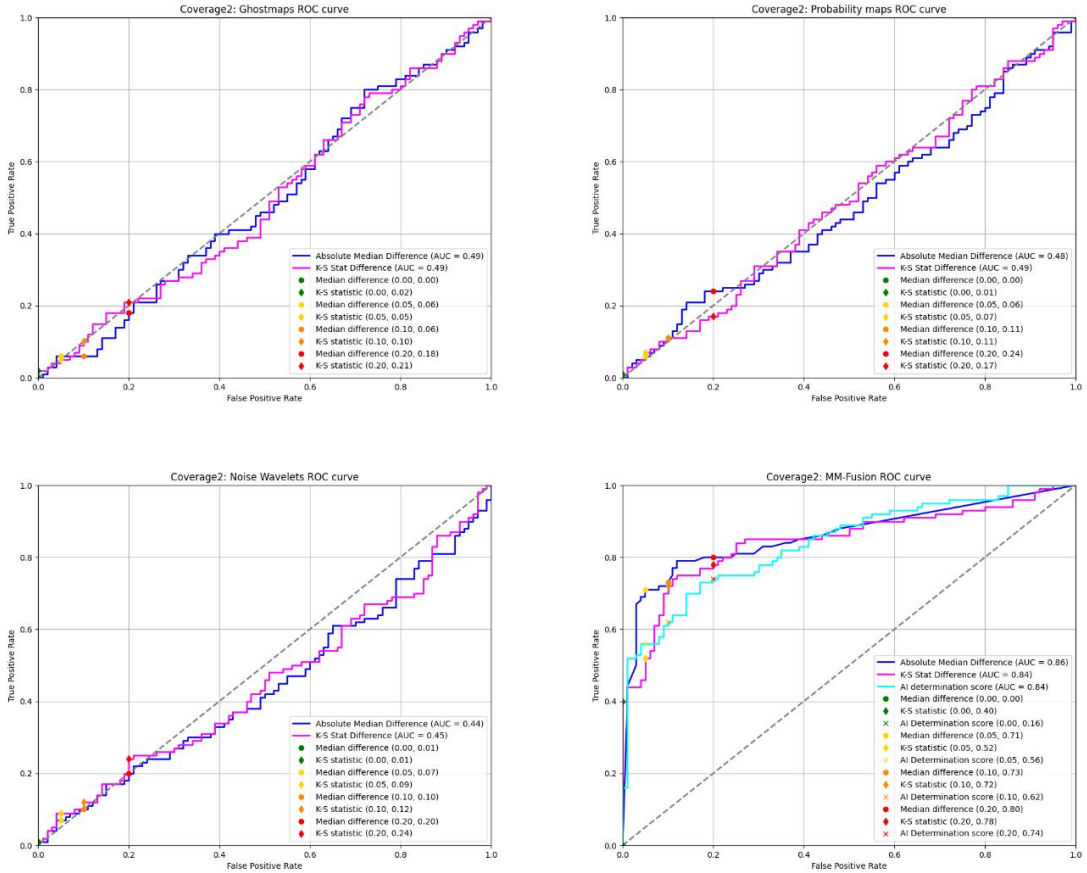

Figure 32: Results for the Coverage dataset when using the copy mask for evaluating original images. The results show that all proposed classical techniques perform at chance or slightly below for detecting copy move artifacts in uncompressed images. This finding demonstrates the unreliability of using random masks for estimating false positives. Consequently, ROC curves for traditional techniques are invalidated for all datasets. Specifically for the Coverage dataset, this is the expected result and indicates the algorithms work correctly and as intended by the scientific principles from which they were developed.

The quantitative results for the Coverage2 dataset show that all of the proposed classical techniques perform at chance or slightly below and thus prove unreliable for detecting copy move artifacts in uncompressed images. This is in line with expectations and is an indication that the algorithms work correctly and as intended by the scientific principles from which they were developed.

The Coverage2 results also demonstrate the unreliability of using random masks for estimating false positives. Thus, the ROC curves for all traditional techniques shown in this work are invalidated. Only the "AI determination score" AUC's for the MM-fusion model remain reliable for interpretation.

The remainder of this work will focus on notable findings and theories based on qualitative investigation of the results from the study. A conclusion comparing classical techniques to state-of-the-art AI networks cannot be drawn until reliable solutions are developed for detecting false positives for each classical technique.

3.5 Qualitative study between AI and traditional techniques

3.5.1 Ghost map performance

When examining some false positive results in the datasets, a number of 'strong' false positives seem to be caused by overexposure or underexposure of the camera in bright or dark conditions, causing a saturated or uniform region. An example is shown in figure 33. From a computer algorithm's perspective, this is a genuine false positive. However, a human analyst, considering the contents of the image should come to the conclusion that the image is plausibly authentic.

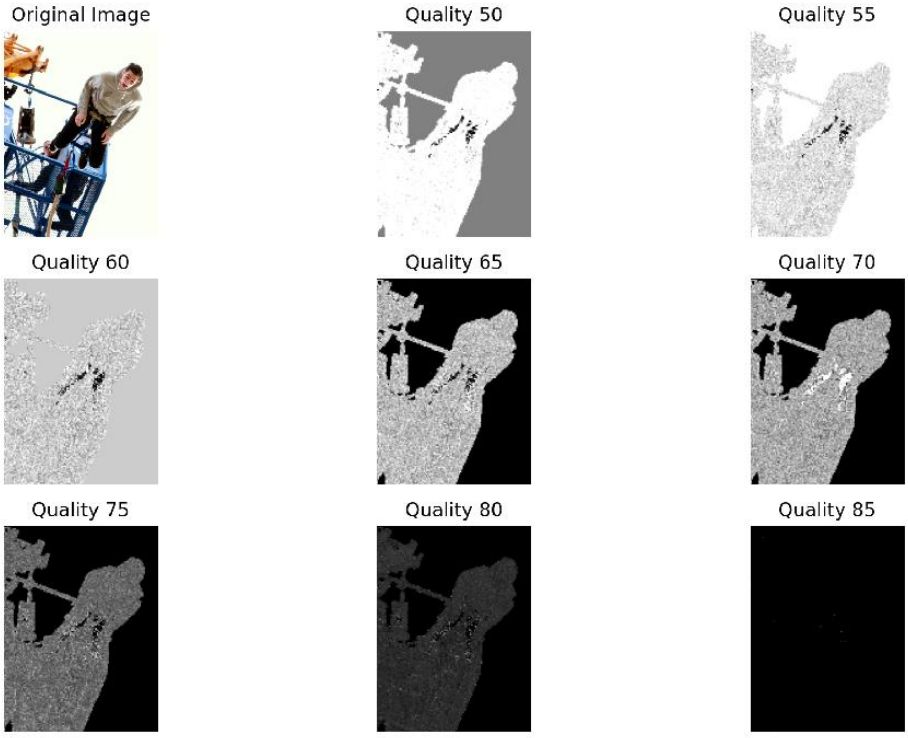

Figure 33: Example of a false positive result caused by overexposure in bright conditions. The sky is uniform which causes this region to appear as manipulated in the image.

Despite the unreliability of using a random mask for detecting false positives in original images, it succeeded in catching a mislabeled manipulated image in the IMD2020 dataset. Figure 34 shows the ghost map output for this image.

For the image, it is evident that the background is uniform and ghost maps confirm this observation. It is unexpected how the background flips from black to white at different compression levels, but the important observation is that the area is uniform in the ghost maps output at all compression levels. When the image was relabeled to manipulated, it increased the detection results at FPR = 0% for ghost maps by 15%. The noise wavelet and probability maps also increased their detection rate by 3% and 8% respectively at FPR = 0% (K-S statistic predictor).

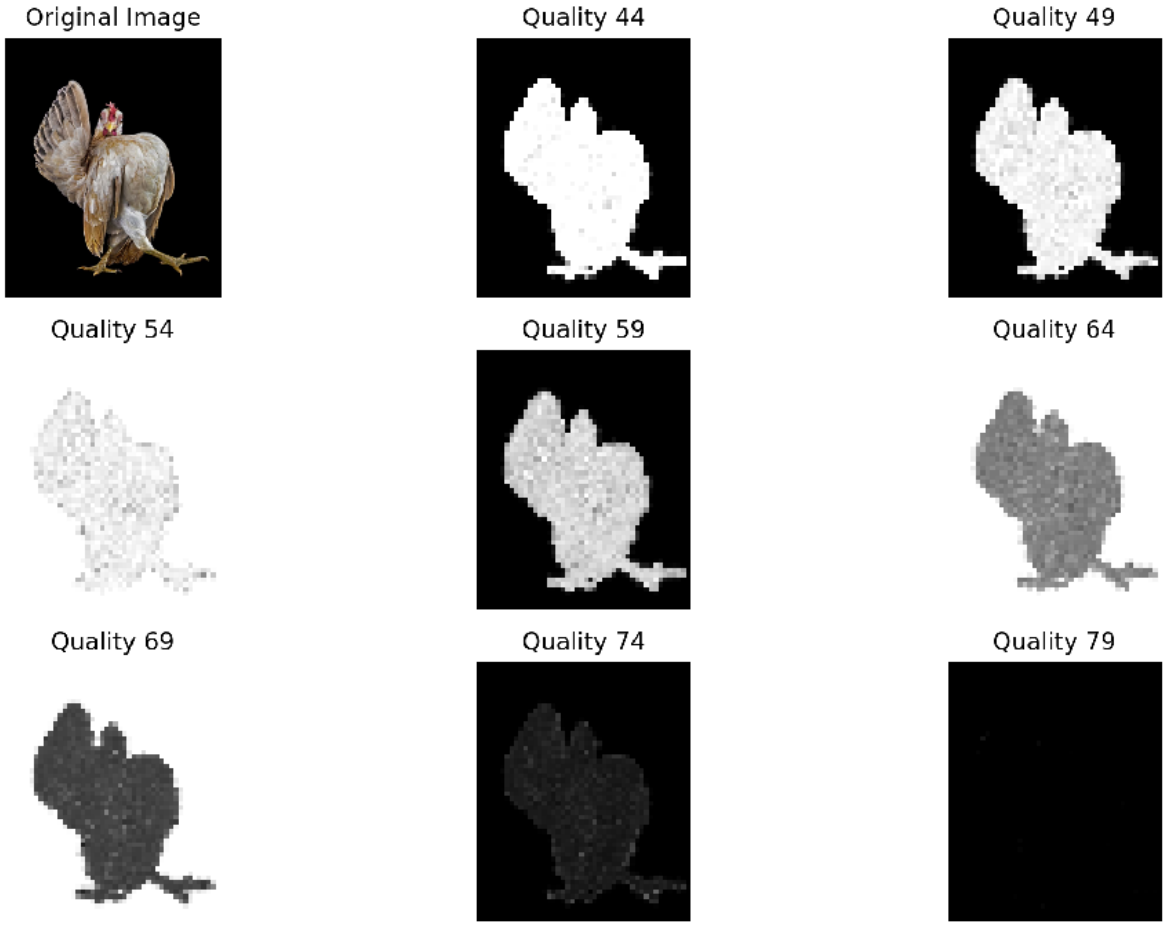

Figure 34: Ghost map output for a mislabeled manipulated image in the IMD2020 dataset. The uniform background is confirmed by ghost maps, which show consistent uniformity across all compression levels.

Another mislabeled image was found during data analysis and shown in figure 35. The detection threshold for this “false positive” was similar to multiple “real” (=plausible) images displaying under- or overexposure.

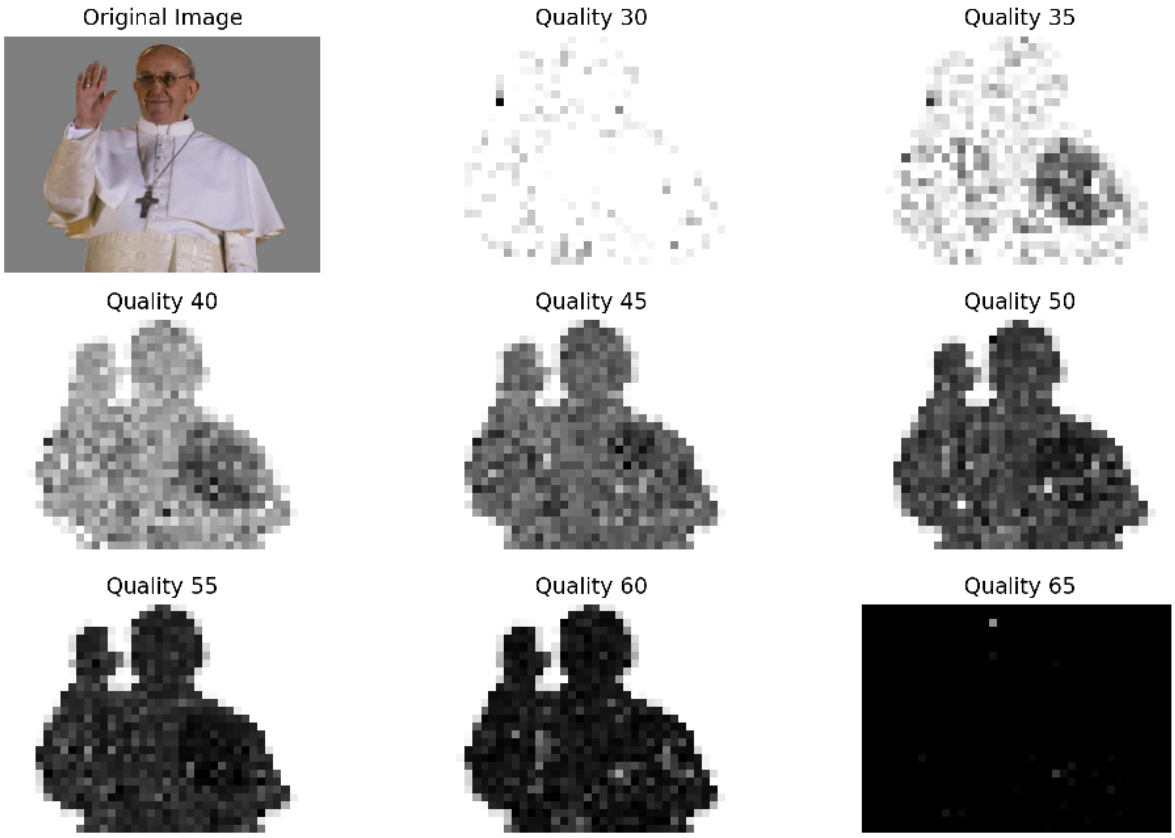

Figure 35: Another example of a mislabeled image found in the IMD2020 dataset. The detection threshold for these images was similar to several plausible images displaying uniform regions, frequently due to lighting conditions.

The detection of manipulated images labeled as original in the IMD2020 dataset raises concerns. Such inclusions can misguide AI learning methods, possibly preventing them from effectively learning characteristics of manipulated images.

Notably, the AI model MM-Fusion was trained on this dataset and failed to properly classify these images as manipulated. This finding highlights the importance of accurate labeling and the potential pitfalls of using flawed datasets for training AI models.

The authors of the IMD2020 dataset have been notified of these findings.

3.5.2 Probability maps performance

Probability maps produced no reliable examples for detecting resampled artifacts. Most flagged detections were due to uniform areas surrounding the manipulated region. Figure 36 shows an example.

Another finding is that the probability maps seem to reliably indicate regions with high noise levels which is an interesting finding. This observation, while intriguing, could have been predicted. Noise, by its nature, disturbs pixel correlation. It logically follows that regions with anomalous levels of noise show weak correlation levels relative to regions without noise. This idea can be confirmed by comparing the probability map with the noise wavelets output map, as shown in the figure 36.

In conclusion, probability maps are not useful for detecting interpolated artifacts since uniform regions and high JPEG compression show similar activation levels. In the pursuit of detecting resampled artifacts, probability maps are only useful when combined with a Fourier map. The

experiment also indicates that probability maps can be used as an analytic tool for analyzing noise levels in an image.

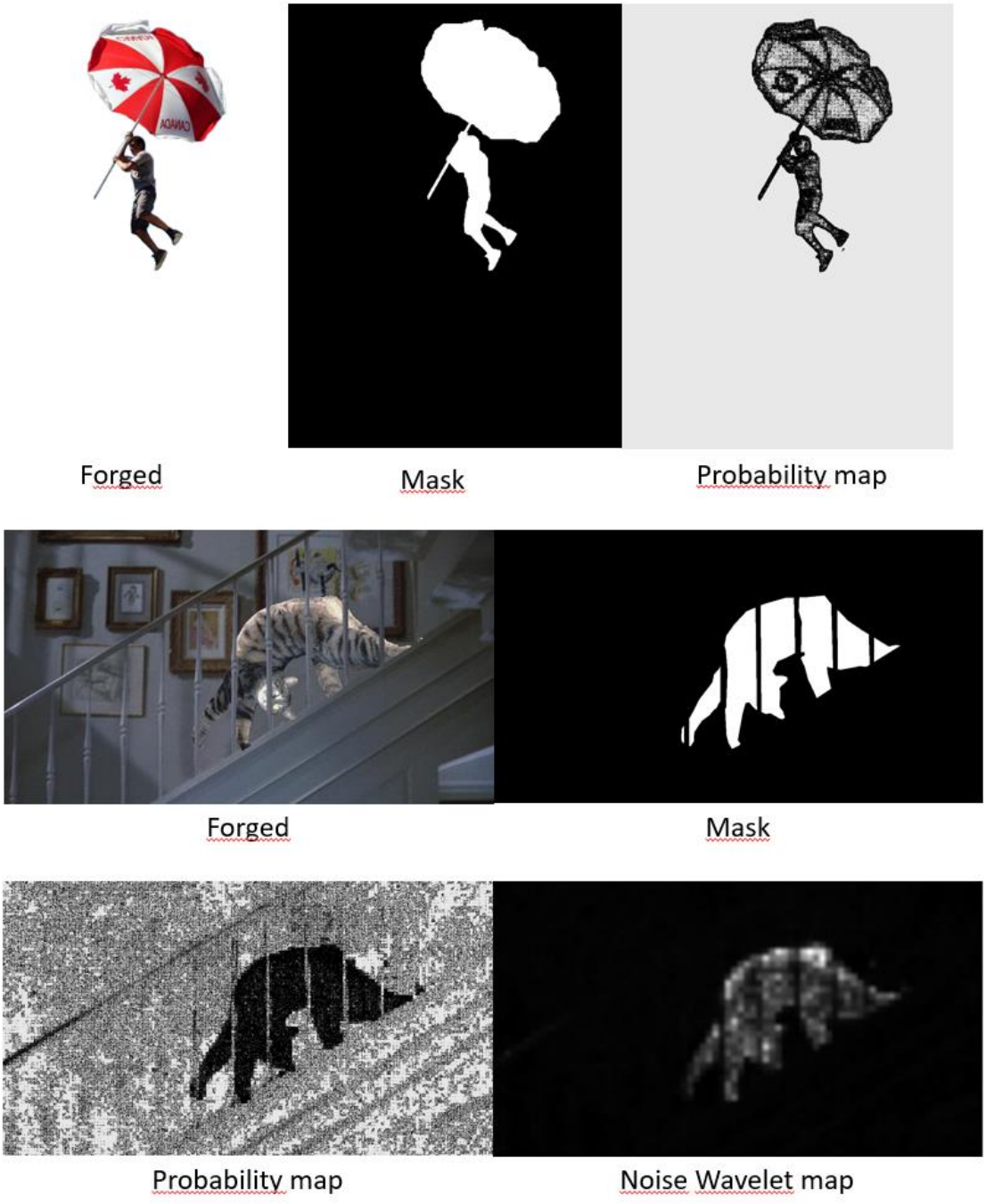

Figure 36: Examples of probability maps being unsuitable for the detection of resampled artifacts. Most flagged detections occurred due to uniform areas surrounding the manipulated region, as shown in the top row example. Probability maps seem to indicate regions with high noise levels. This indication is supported by comparing the probability map with the noise wavelets output map (row 3). This highlights that while probability maps are not useful for detecting interpolated artifacts due to similar activation levels in uniform regions and high JPEG compression, they can serve as analytic tools for analyzing noise levels in an image.

3.5.3 Noise Wavelets performance

Remember that noise wavelets are recommended to be used in conjunction with other techniques, as noise patterns can naturally occur in authentic images. A typical detection example is shown in figure 37.

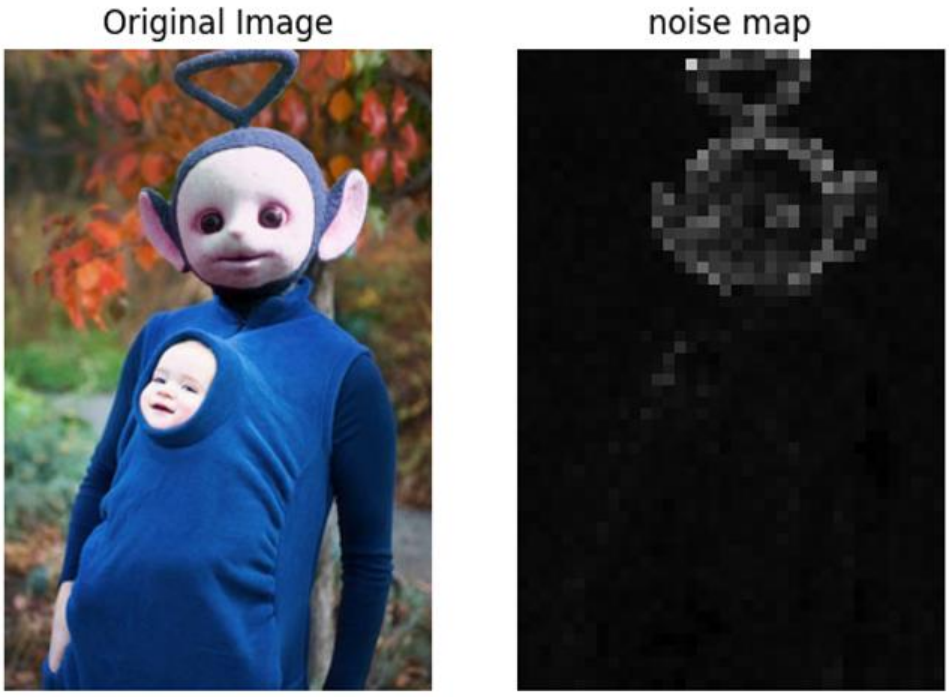

Figure 37: Typical detection example using noise wavelets. The head of the person stands out as a clear anomaly.

Noise can appear randomly, but editing an image seems to alter noise levels depending on the manipulation or algorithm used. This observation is not unusual since software generally alters pixel contents in a deterministic manner, potentially disrupting noise artifacts present in authentic images.

Figure 38 shows two original images and a combined splice from the Columbia dataset. The manipulated region appears as a large anomaly in the splice output map. This is rather interesting because the same region shows different activation levels in its original picture. The two following theories could explain this phenomenon:

1. Algorithm-Induced Noise Alteration: The algorithm used to create the splice may have altered the noise patterns in the original images, but their distinct characteristics remain visible in the splice.
2. Relative Comparison Effects: The noise wavelet algorithm rescales the noise spectrum from 0 to 255 for better visualization. As a result, we perceive a full spectrum for each original image. Combining these images together causes the original noise intensities to be rescaled relative to each other. This allows the detection of anomalies if the difference in original noise characteristics are sufficiently large.

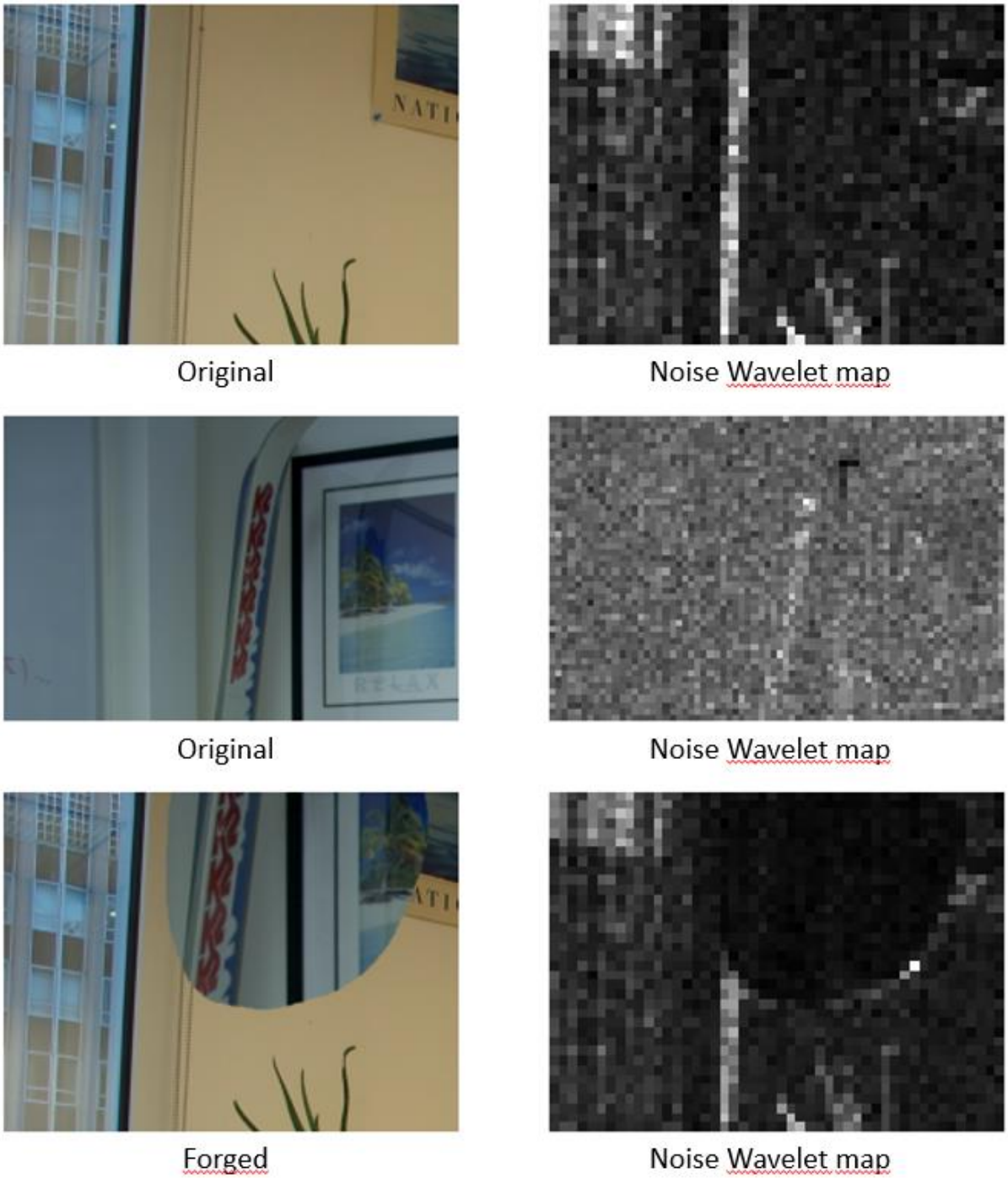

Figure 38: Two original images and a combined splice from the Columbia dataset, demonstrating how the manipulated region appears as a large anomaly in the splice output map. This phenomenon could be due to algorithm-induced noise alteration or relative comparison effects, because the noise in one image could be comparatively weak to noise in the other image.

The images in figure 39 were the strongest false positives for the noise wavelet algorithm in the IMD2020 dataset. Interestingly, these images were also the strongest false positives for the probability maps algorithm. This further solidifies the findings that probability maps can be used as a tool for noise investigation. The phenomena of zero noise in the backgrounds of these images could potentially be explained by camera lens focus, which somewhat blurs the background, resulting in the absence of noise artifacts.

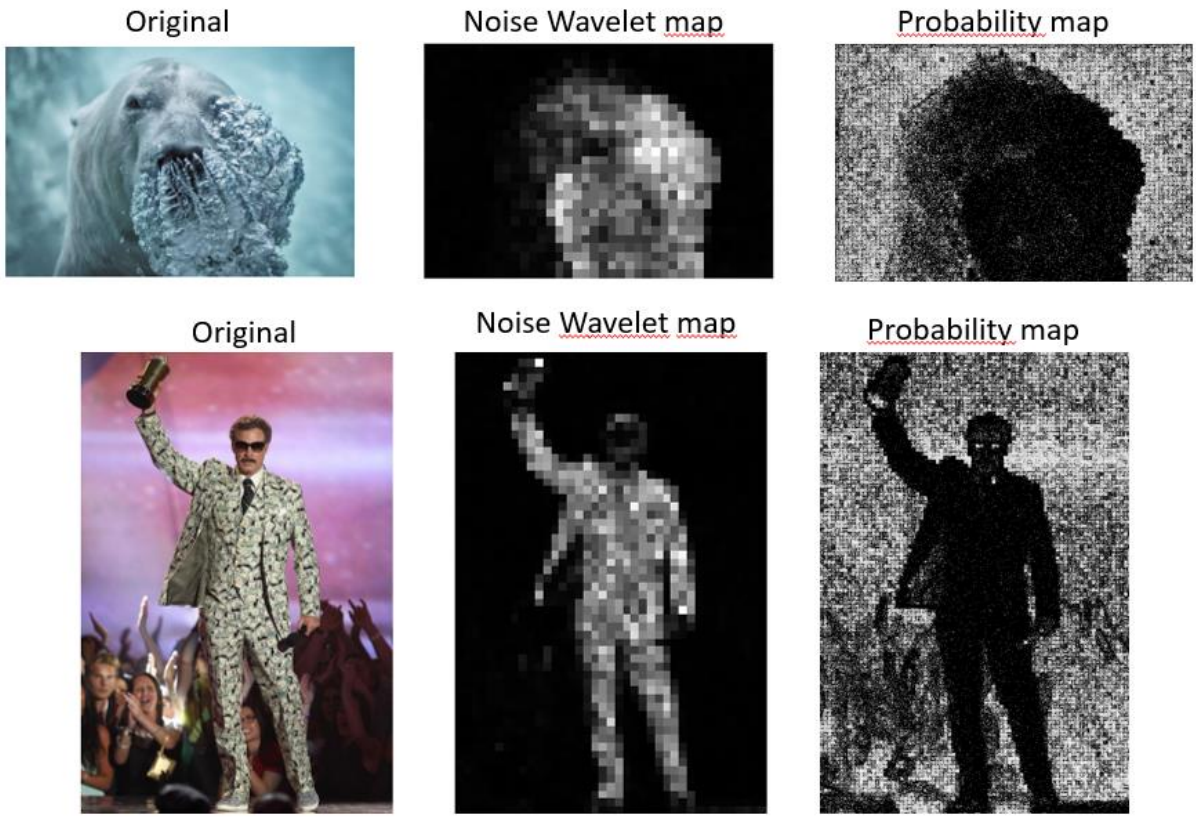

Figure 39: The strongest false positives for the noise wavelet and the probability maps algorithm in the IMD2020 dataset. This finding further supports the use of probability maps as a tool for noise level investigation. The phenomena of zero noise in the backgrounds of these images could potentially be explained by camera lens focus, which blurs the background, resulting in the absence of noise artifacts.

3.5.4 MM-fusion performance

MM-Fusion demonstrated a strong detection score for an image labeled as original in the IMD2020 dataset. When analyzing this image with other techniques, it was proven that MM-Fusion is correct. Figure 40 shows an analysis where the ghost map algorithm proves this image was manipulated.

In the example image, it is visually apparent that something on the blackboard to the left of the laptop was blotted out. MM-Fusion correctly highlights this manipulated region and the same region becomes highly salient in the ghost maps at quality 75.

Figure 40 confirms that quality 75 is in fact a ghost that reveals itself when the JPEG grid is correctly aligned. This observation allows us to dismiss other strange artifacts at different qualities such as quality 65 & 70. These artifacts are likely caused by noise or uniform areas caused by dim lighting.

Comparing these offsets with the original JPEG ghosts, it seems that the laptop screen was also intentionally blotted out. First, the screen is suspiciously uniform and it is unlikely for a laptop to emit enough light to oversaturate a camera when other light sources are also present in the room. Additionally, a suspicious gray box has formed around the laptop screen at quality 75 where the ghost reveals itself. This is similar to figure 5 where the center part of the image demonstrated a gray box structure because the image was saved at a higher quality, which created composite JPEG blocks that manifested in this manner. The quality estimation in figure 41 reveals that the image was indeed resaved at a higher quality, adding extra substance to this theory, proven by the presence of a JPEG ghost near quality 75.

Because this artifact is not highly salient and it manifests around uniform regions, the evidence for this theory might not be reliable. A qualitative study to specifically investigate this phenomenon in various circumstances is needed to determine its validity.

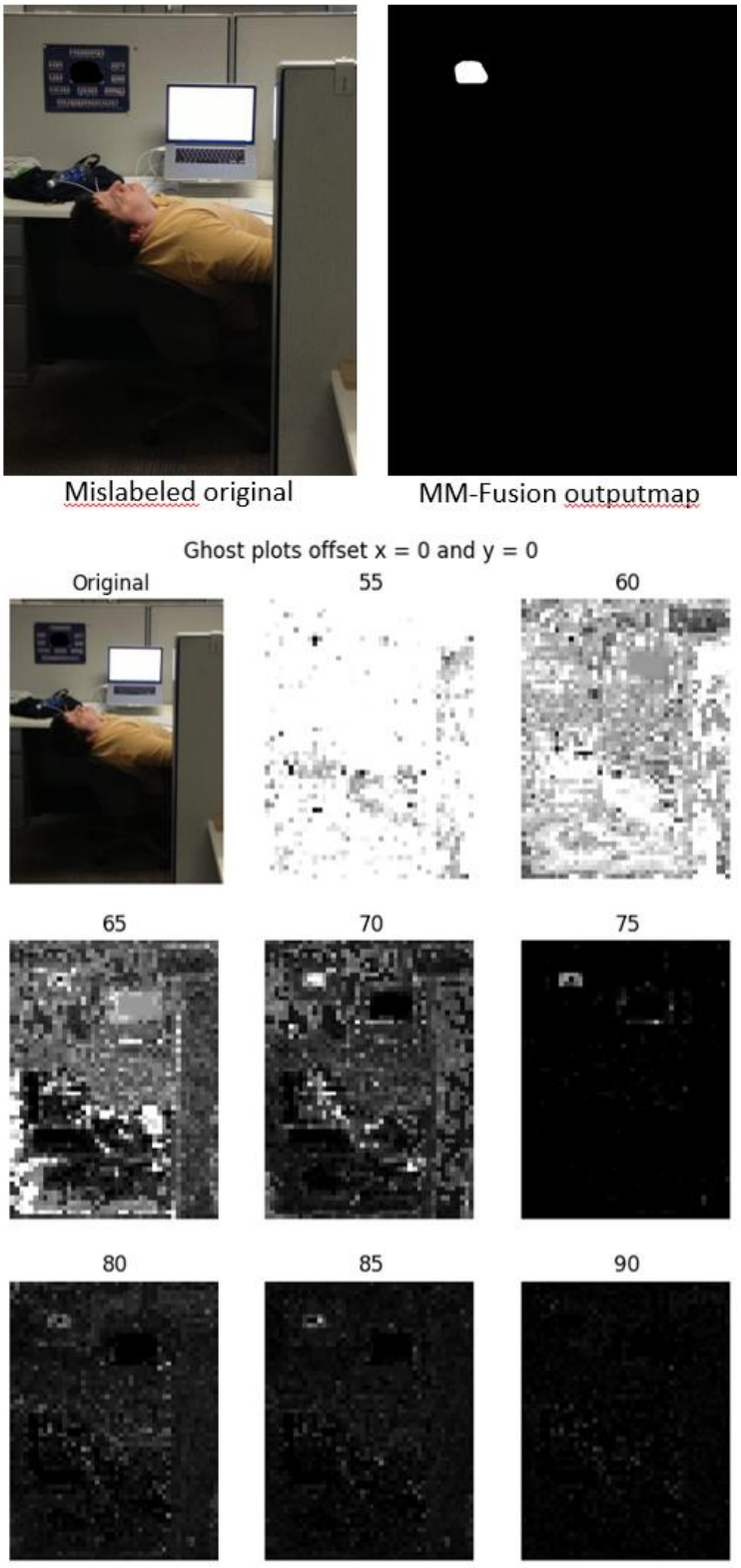

Figure 40: Analysis of a mislabeled original image from the IMD2020 dataset, where MM-Fusion correctly identified the manipulation. The ghost map algorithm confirms the manipulation, highlighting at JPEG quality 75 that a region on the blackboard to the left of the laptop that was blotted out.

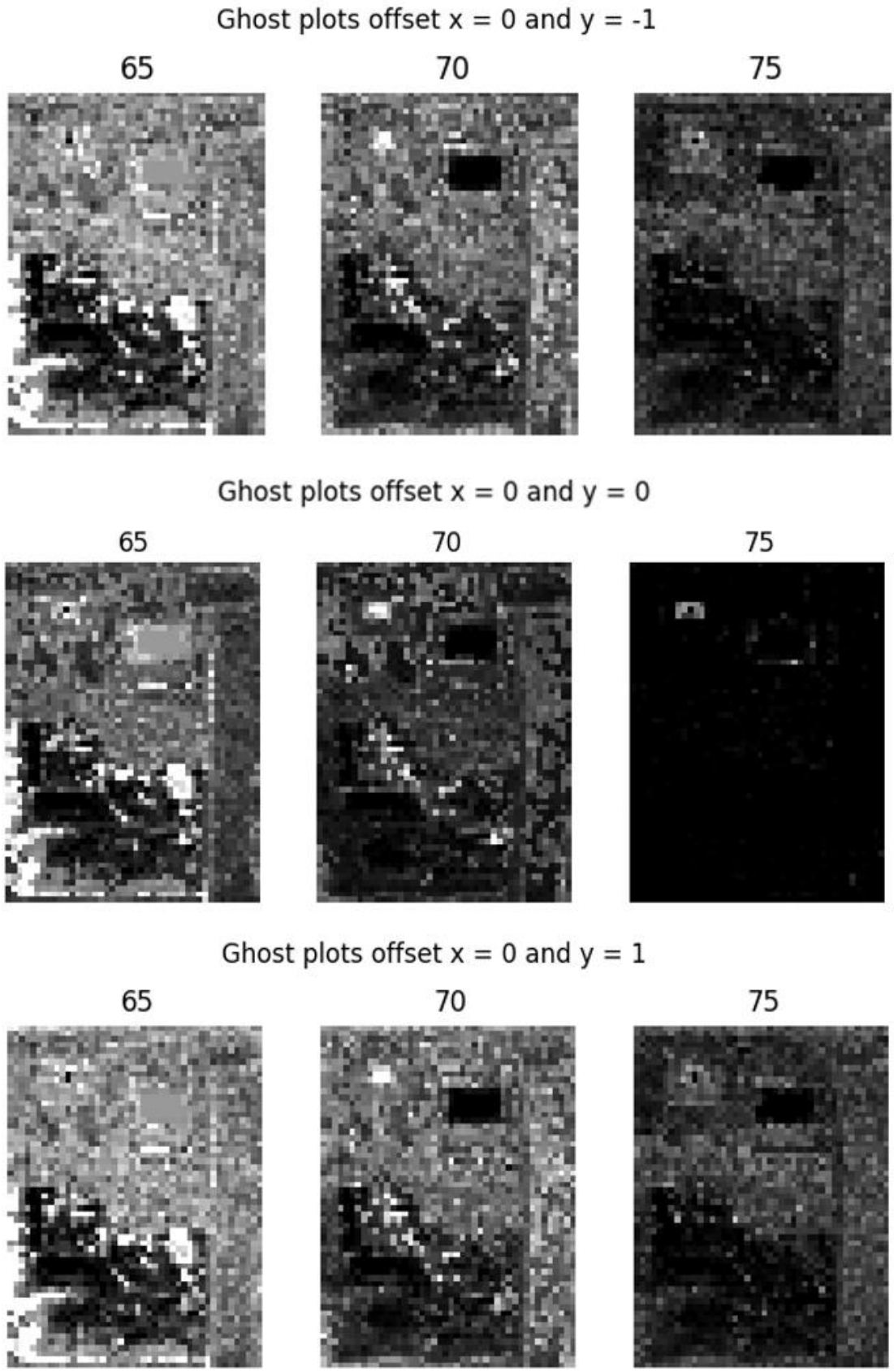

Figure 41: Ghost map outputs at three neighboring JPEG offsets (y = -1, y = 0, y = 1). For the correct offset y = 1 a JPEG ghost reveals itself at quality 75. This observation allows dismissal of other artifacts at different qualities, such as 65 and 70, which are likely caused by noise or uniform areas due to dim lighting.

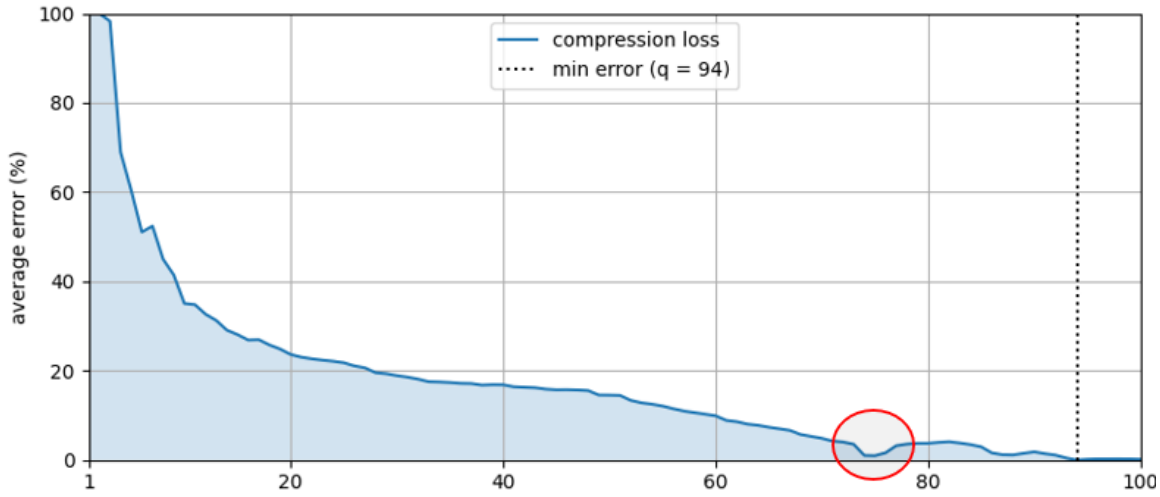

Figure 42: Quality estimation shows the image was resaved at a higher quality by showing the presence of a JPEG ghost near quality 75.

3.5.5 Direct comparison of detections between MM-Fusion and traditional methods

In this section, we present a comparative analysis between the state-of-the-art MM-Fusion model and traditional techniques for detecting and localizing manipulated regions in digital images. We will compare the output of the MM-Fusion model to traditional techniques for multiple examples shown previously in this work.

Figure 43 shows the MM-Fusion output for a manipulated image of a cat, compared to the output from the Ghost maps technique. The Ghost maps technique successfully localizes all manipulated areas, whereas MM-Fusion fails to achieve the same accuracy, missing the manipulated area around the left paw.

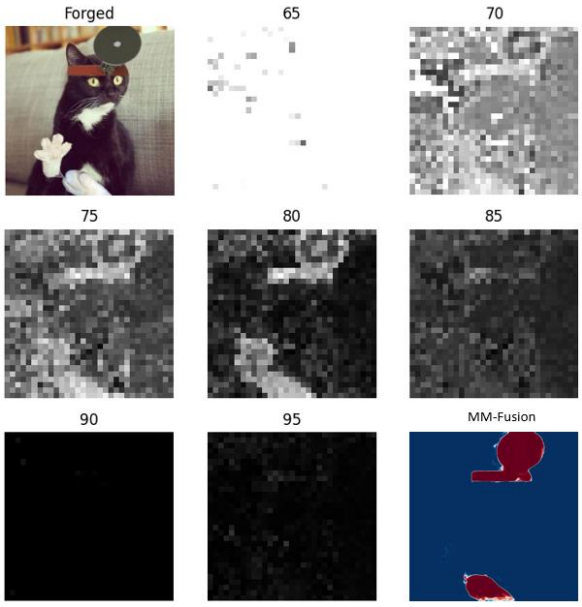

Figure 43: Comparison of MM-Fusion and Ghost Map Outputs for a manipulated image of a cat. The ghost map at quality 80 correctly identifies all manipulated areas in the image, while MM-Fusion fails to the manipulated region for the left paw. This demonstrates a limitation in MM-Fusion's localization accuracy despite the fact that these manipulations are the result of the same JPEG compression artifact.

Figure 44 presents an example where an additional manipulation was introduced into the manipulated cat image. Specifically, a region was copied and the image was saved at its original JPEG quality of 90. The results from ghost maps proof consistent and predictable, successfully detecting the new manipulation alongside previous manipulations. In contrast, MM-Fusion successfully identifies the new copied region, but fails to detect the bottom paw which it previously detected. This discrepancy is an example of an adversarial attack, which as discussed in this work, is a vulnerability of AI networks according to many studies. In a blind test, one might trust the AI model based on its reported statistical accuracy of 0.897 AUC; however, this example demonstrates that reliance on such statistics can be misleading. While the missed manipulation may be inconsequential in this case, in higher-stakes scenarios, this behavior is potentially a significant drawback. To further illustrate this potential downside, consider a case where the ghost maps algorithm fails to detect a manipulation. The conclusion then would be that no differences in JPEG compression rates were observed in the image. However, when the AI model fails to detect a manipulated region, one might erroneously conclude: "With a high degree of certainty, this area is authentic.". Figure 45 shows another example of an adversarial attack, where a copied region was added to the image.

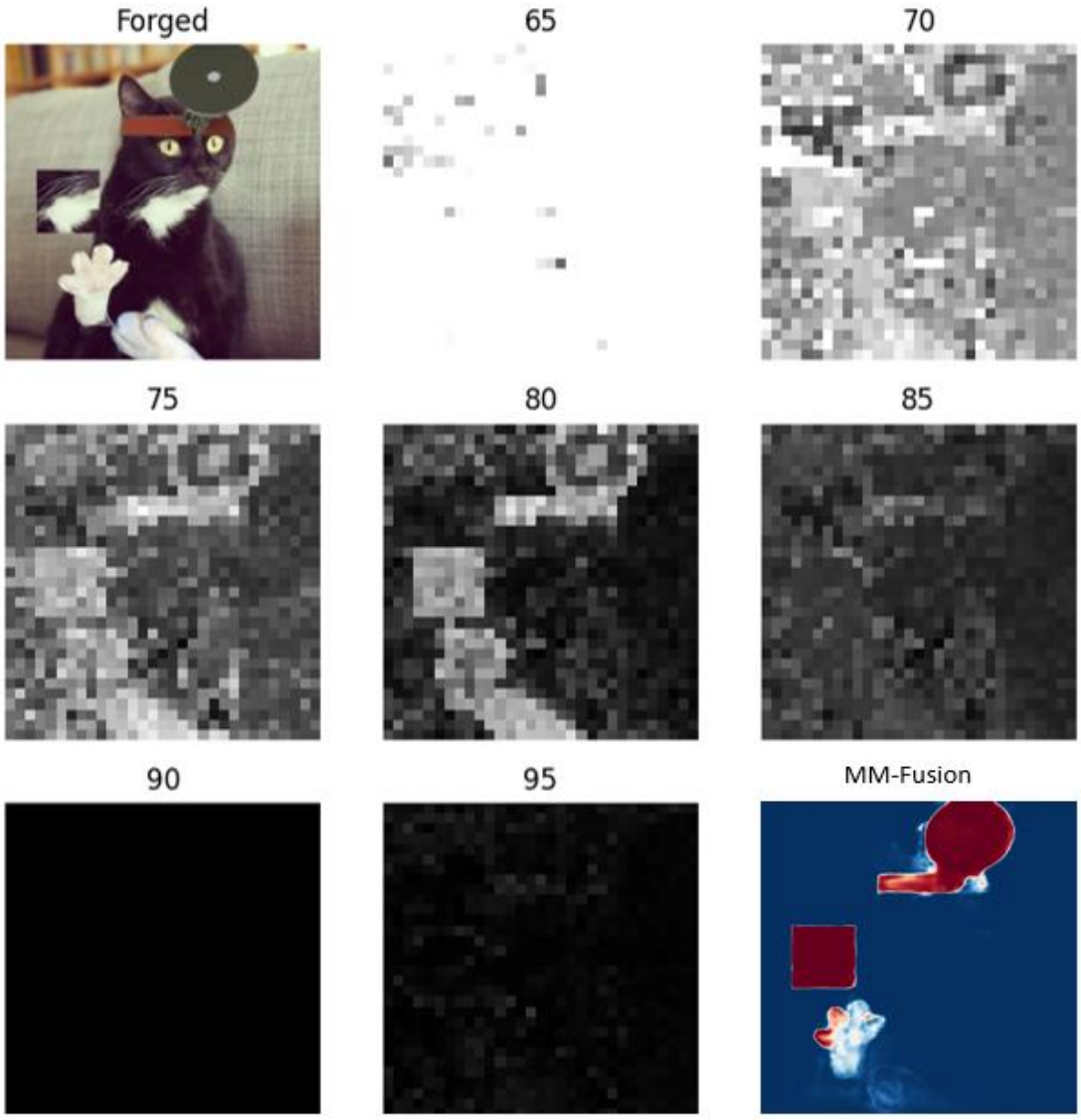

Figure 44: Demonstration of an adversarial attack on MM-Fusion. A new copied region was introduced to the manipulated cat image and saved at its original JPEG quality of 90. The ghost map method reliably detects both the original and new manipulations, whereas MM-Fusion fails to detect manipulations which it previously did correctly identify.

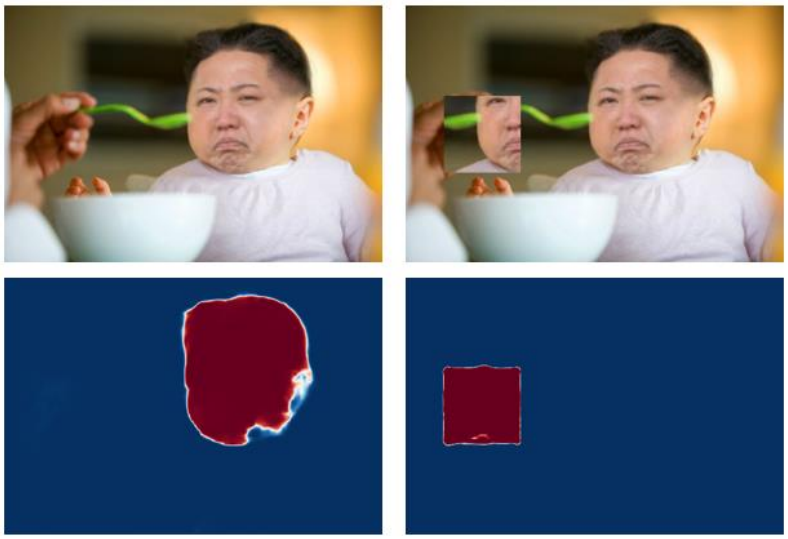

Figure 45: Demonstration of the same adversarial attack for a new example. A copied region was inserted (top right), which affects MM-Fusion's detection capabilities.

Figure 46 presents the MM-Fusion output for two earlier examples in this work, where resampling analysis definitively identifies the manipulated regions in both examples (figure 10 and 11). While MM-Fusion correctly highlights the manipulations in higher-quality JPEG images, it fails to do so for more heavily compressed images.

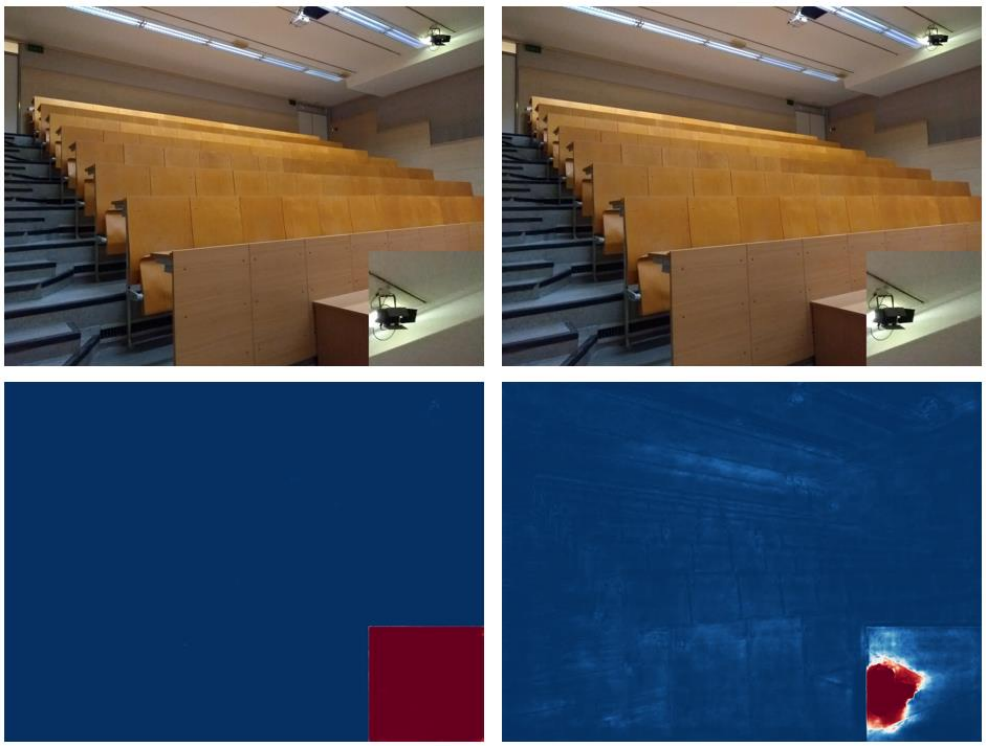

Figure 46:left: Manipulated image at JPEG quality 100%. Right: The same manipulated image at quality 60. Demonstrating the falling localization accuracy for MM-Fusion under heavy JPEG compression. Both images were previously examined with a resampling analysis in figure 14 and 16 respectively.

In conclusion, these examples illustrate the continued vulnerability of state-of-the-art AI models to adversarial attacks. Furthermore, these examples were readily identified when comparing a few example outputs. These findings suggest that caution must be exercised when utilizing AI driven approaches in the field of digital image forensics, despite their statistical performance. They also indicate the continued relevance of traditional techniques and the need for consistent, transparent and explainable results which are crucial for making informed decisions.

3.6 Computational cost of algorithms

The computational cost of an algorithm can determine its suitability for a specific task. For instance, while computational cost may be negligible when analyzing a few images for a legal case, it can become prohibitively large when screening large volumes of user-uploaded content to prevent misinformation on a website.

In this study, we evaluate the computational cost by averaging the time required to calculate an output map for each algorithm across different image resolutions. This approach provides a clear understanding of how computational cost scales with image size and allows for practical insights that are transferable to different projects.

Our methodology was as follows:

First, 10 images at various resolutions were selected from the different datasets used in this study. Then the output maps at various resolutions were calculated and the processing time per image was recorded. This isolation ensures that the average time reflects only the cost of generating the output map, excluding other processing operations such as evaluating the image using a K-S statistic or other evaluation methodologies. Specifically for the ghost maps algorithm, the parameters were set to calculate the ghost maps from JPEG quality 60 to 100 in steps of 5, calculating a total of 9 output maps per image.

Important system specifications are an AMD Ryzen 7 5700x processor, 16GB of RAM, and an Asus Geforce RTX-3070 TUF graphics card. No particular steps were taken to improve computation times.

Table 3: Computational Cost of Algorithms showing the average time required in seconds to calculate the output maps for each algorithm across different image resolutions.

	256x256	568x757	768x1024	1200x1600	1536x2048	2304x3072	3240x4320
Ghost	0,042	0,329	0,671	1,502	2,408	5,350	10,566
Probability	1,813	11,347	9,631	27,627	46,205	98,563	172,534
Noise	0,004	0,011	0,027	0,074	0,159	0,310	0,549
MM-Fusion	0,234	0,382	0,949	1,422	5,348	4,654	4,984

The computational times listed in Table above reveal some interesting insights.

1. **Ghost maps**: Exhibits a linear trend per pixel increase. Detailed analysis showed that calculation times per image varied by less than a second at each resolution.
2. **Probability Algorithm**: Does not exhibit a linear trend per pixel increase. Detailed analysis showed that calculation time per image can vary depending on image content. For all resolution, some images took 2-4 times longer to process than others. For example, at resolution 3240x4320, processing times were between 140 to 350 seconds per image.

3. **Noise Algorithm**: Maintains a sub second computational cost across all resolutions. An average time of 0,498s the highest resolution indicates its potential for real-time applications.
4. **MM-Fusion Algorithm**: Maintains reasonably low computational costs at all resolutions. The decrease in computational cost from resolution 1536x2048 to 2304x3074 suggests the algorithm's performance varies based on factors beyond resolution. Since MM-Fusion is an AI model, the different computational costs per resolution could be partially influenced by the data on which it was trained.

These findings can help practitioners in making informed decisions when selecting algorithms depending on the computational constraints of their application.

4. Future vision

In this section, we explore potential directions for future efforts in the field of digital image forensics. Following our open source philosophy, the addition of more techniques to Sherloq will always be beneficial. Therefore we will focus particularly on the role of AI and its integration with traditional forensic techniques.

4.1 AI in a Key Support Role

The issue of overfitting in AI-driven approaches has been highlighted as a drawback multiple times in this work. However, rather than viewing overfitting as a limitation, it can potentially be leveraged as a significant advantage, depending on the application.

Our vision for the future of digital image forensics involves the development of an AI-supported toolbox. This toolbox would include highly specialized AI networks, each trained to detect specific types of manipulations, which can guide practitioners and traditional algorithms. A prerequisite for this approach would be the construction of a large, trusted dataset of authentic images, containing many different camera models, all types of scenery and various compression rates.

For example, traditional techniques struggle with copy-move detection tasks due to the nearly infinite search space. To address this, an AI could be trained on millions of examples generated by a script that created random copy-move forgeries from images in the trusted authentic dataset. The script would generate thousands of random examples for various scenarios of copy-move operations, such as rotations, resizing and color changes. Once trained, this AI could quickly and accurately identify potential copied regions, allowing traditional algorithms to focus their computational resources more effectively.

This process can be repeated for JPEG artifacts, interpolation artifacts and noise artifacts. A benefit of this approach is that AI is generally faster at analyzing images compared to many traditional algorithms, especially for large images, as discussed in the computational cost section of this work. Forensic analysts could employ these AI networks to rapidly highlight potential areas of interest and also identify which techniques are most likely to yield a successful analysis. Analysts would then only need to verify these results using traditional methods.

4.1 The potential of Explainable Artificial intelligence

To address the “black-box” character of AI, the research field of Explainable Artificial intelligence (XAI) has emerged and is rapidly evolving [73]. Despite considerable efforts, the field is still in its infancy and faces many challenges [74]. However, if AI models become sufficiently explainable in the future, they could provide critical insights that lead to the development of new traditional techniques.

For example, the work by Gonzalo et al. [46] demonstrated how AI can be trained to detect generated human faces with remarkable specificity. Their network was so specialized that it only detected generated images containing human faces, ignoring other types of generated content. This finding suggests that the network may have latched onto a particular artifact unique to most generated faces. If advancements in XAI allow us to sufficiently understand this artifact, the insights gained could be formalized into a new detection algorithm that does not rely on AI. This approach has the potential to significantly enhance the robustness and reliability of digital image forensics.

5. Conclusion

The rapid advancement of artificial intelligence and image-editing software has fundamentally transformed the landscape of digital imagery, challenging long-held beliefs about the reliability of visual evidence. In this context, the field of Digital Image Forensics has emerged as a crucial discipline, providing the tools and methodologies necessary to discern authentic images from those that have been manipulated. This thesis has sought to contribute to this field by performing a quantitative and qualitative comparison between the combined performance of traditional forensic techniques and a recent state-of-the-art AI method.

Another central aspect of this work was the development and integration of three algorithms into the Sherloq open-source image forensic toolset. Each algorithm was chosen because it exploits a different type of artifact and because it was novel to Sherloq. JPEG artifacts were exploited with ghost maps, a technique pioneered by Hany Farid; resampling artifacts were exploited with Probability Maps combined with Fourier transforms, a technique pioneered by Popescu et al.; and noise artifacts were exploited with Noise Wavelet analysis, a technique pioneered by Mahdian et al. Each algorithm was rigorously tested and validated, with their implementation demonstrated to be loyal to the scientific principles on which they are based and capable of reproducing the findings demonstrated in their pioneering works. In line with the open-source philosophy that underpins this work, the algorithms together with user interfaces that facilitates their use, were integrated into Sherloq. These contributions received positive response from the community and will hopefully inspire further collaboration and development.

The quantitative comparison revealed that the combined performance of the ghost maps, Probability Maps, and Noise Wavelet algorithms surpassed that of the state-of-the-art MM-Fusion model created by Triaridis et al. by an average of 17% more detections in scenarios where a 0% false positive rate was allowed. This suggests that, despite the impressive capabilities of AI, traditional methods remain relevant, especially in high-stakes environments where the accuracy and explainability of results are paramount.

This study also explored the effectiveness of the K-S statistic compared to the absolute median difference for classifying images as either manipulated or authentic. The K-S statistic consistently outperformed the absolute median difference in the quantitative tests for all datasets and algorithms, with an AUC that was on average 9,42% higher. Further analysis of detection overlap revealed that at every False Positive Rate (FPR) level, the K-S statistic overlapped with the absolute median difference by 60% to 100% in terms of correct detections across all methods and datasets. A theoretical and practical example was also presented to demonstrate that the absolute median difference is more susceptible to specific pixel distributions and can be significantly influenced by a small number of pixels. Conversely, the K-S statistic proved to be more robust and reliable in indicating similarity between two samples.

Additionally, we examined the utility of Probability Maps in detecting resampling artifacts without accompanying Fourier transforms. The results indicated that Probability Maps alone are not reliable for this purpose, as uniform regions and high compression levels also produce high activation in Probability Maps, making it difficult to distinguish these from resampled artifacts without Fourier analysis. However, Probability Maps were found to be a reliable indicator for noise analysis, which is plausible given that noise disrupts pixel correlations when are detected by the probability map. We provided examples where Noise Wavelet analysis and Probability Maps yielded consistent conclusions.

Another notable finding was the strong performance of Ghost Maps on the IMD2020 and In the Wild datasets. These datasets simulate real-world scenarios for how manipulated images are encountered on the internet, making the success of the Ghost Map algorithm particularly significant. Although offset checks were not performed in this study, it would be intriguing to study how well Ghost Maps would perform when all offsets are considered. The preliminary results are promising and warrant further investigation.

The qualitative comparison further supported findings suggested by the quantitative study, highlighting instances where the MM-Fusion model failed to detect manipulations that were effectively identified by the traditional algorithms and also demonstrated the vulnerability of MM-Fusion to adversarial attacks. These findings challenge the notion that AI can replace traditional forensic methods.

However, it is important to acknowledge the limitations of this study. Biases were identified in the validation process using masks, particularly in the evaluation of authentic images that do not have a mask. Future work should refine the validation process of authentic images in order to validate the findings of the quantitative comparison. Suggestions on how this validation process can be improved is addressed in this work.

This work also puts forward a vision for future research and developments in the field. One of the key areas for future exploration is the integration of AI in a support role within the forensic analysis process. As discussed, overfitting—a commonly cited drawback of AI—has the potential to be harnessed to create highly specialized networks capable of detecting specific types of manipulations. These AI networks could serve as an initial filter, quickly identifying areas of interest within an image, thereby allowing traditional techniques to be applied more efficiently and effectively. Moreover, the potential of Explainable Artificial Intelligence (XAI) represents an exciting frontier for digital forensics. If AI models can be made sufficiently transparent, the insights gained from their decision-making processes could lead to the development of new traditional techniques.

In summary, this thesis has contributed to the field of Digital Image Forensics by developing and validating the functionality of a set of traditional forensic tools and adding them to the public domain, conducting a comparative study between these algorithms and a state-of-the-art AI model, and sharing a vision for the future of forensic analysis. While AI presents exciting possibilities, this research reaffirms the importance of traditional forensic techniques and suggests that a balanced approach that leverages the strengths of both AI and classical methods, can play an important role in addressing the challenges of image authentication in the digital age.

References

Bronnen

[1] Farid, H. (2022). Creating, Using, Misusing, and Detecting Deep Fakes. Journal of Online Trust & Safety, 1(4). https://doi.org/10.54501/jots.v1i4.56

[2] Sofer, C., Dotsch, R., Wigboldus, D. H. J., & Todorov, A. (2015). What Is Typical Is Good: The Influence of Face Typicality on Perceived Trustworthiness. Psychological Science, 26(1), 39–47. https://doi.org/10.1177/0956797614554955

[3] 3. Shen, B., Webster, B. R., O'Toole, A., Bowyer, K. W., & Scheirer, W. J. (2021). A study of the human perception of synthetic faces. In 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021) (pp. 1-8). IEEE. https://doi.org/10.1109/FG52635.2021.9667066

[4] Tucciarelli, R., Vehar, N., Chandaria, S., & Tsakiris, M. (2022). On the realness of people who do not exist: The social processing of artificial faces. iScience, 25(12), 105441–105441. https://doi.org/10.1016/j.isci.2022.105441

[5] Damiani, J. (2019, September 3). A voice deepfake was used to scam a CEO out of $243,000. Forbes. https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/

[6] 6. Elgin, B. (2020, July 14). Fraudsters use AI to mimic CEO's voice in unusual cybercrime case. The Wall Street Journal. https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402

[7] Ponton, J. L., Yun, H., Aristidou, A., Andujar, C., & Pelechano, N. (2023). SparsePoser: Real-time Full-body Motion Reconstruction from Sparse Data. ACM Transactions on Graphics, 43(1), 1–14. https://doi.org/10.1145/3625264

[8] Jiang, Y., Ye, Y., Gopinath, D., Won, J., Winkler, A. W., & Liu, C. K. (2022). Transformer Inertial Poser: Real-time Human Motion Reconstruction from Sparse IMUs with Simultaneous Terrain Generation. Proceedings - SIGGRAPH Asia 2022 Conference Papers, 1–9. https://doi.org/10.1145/3550469.3555428

[9] How AI influencers are taking over social media. (n.d.). Digital Agency Network. https://digitalagencynetwork.com/how-ai-influencers-are-taking-over-social-media/

[10] AI influencers on Instagram. (n.d.). IZEA. https://izea.com/resources/ai-influencers-on-instagram/

[11] The rise of AI influencers and AI OnlyFans models. (n.d.). SKIMAI. https://skimai.com/the-rise-of-ai-influencers-and-ai-onlyfans-models/

[12] Guo, H., Hu, S., Wang, X., Ming-Ching, C., & Lyu, S. (2022). Eyes Tell All: Irregular Pupil Shapes Reveal GAN-generated Faces. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2109.00162

[13] Hu, S., Li, Y., & Lyu, S. (2020). Exposing GAN-generated Faces Using Inconsistent Corneal Specular Highlights. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2009.11924

[14] Nishino, K., & Nayar, S. K. (2004). The world in an eye [eye image interpretation]. In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004) (Vol. 1, pp. I–I). https://doi.org/10.1109/CVPR.2004.1315066

[15] Yang, X., Li, Y., Qi, H., & Lyu, S. (2019). Exposing GAN-synthesized Faces Using Landmark Locations. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1904.00167

[16] Wang, X., Guo, H., Hu, S., Ming-Ching, C., & Lyu, S. (2023). GAN-generated Faces Detection: A Survey and New Perspectives. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2202.07145

[17] Nightingale, S., & Farid, H. (2022). Synthetic Faces Are More Trustworthy Than Real Faces. Journal of Vision (Charlottesville, Va.), 22(14), 3068-. https://doi.org/10.1167/jov.22.14.3068

[18] Ramadhani, K. N., & Munir, R. (2020). A Comparative Study of Deepfake Video Detection Method. 2020 3rd International Conference on Information and Communications Technology (ICOIACT), 394–399. https://doi.org/10.1109/ICOIACT50329.2020.9331963

[19] Patel, Y., Tanwar, S., Gupta, R., Bhattacharya, P., Davidson, I. E., Nyameko, R., Aluvala, S., & Vimal, V. (2023). Deepfake Generation and Detection: Case Study and Challenges. IEEE Access, 11, 1–1. https://doi.org/10.1109/ACCESS.2023.3342107

[20] Wang, J., Li, Z., Zhang, C., Chen, J., Wu, Z., Davis, L. S., & Yu-Gang, J. (2022). Fighting Malicious Media Data: A Survey on Tampering Detection and Deepfake Detection. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2212.05667

[21] Patil, K., Kale, S., Dhokey, J., & Gulhane, A. (2023). Deepfake Detection using Biological Features: A Survey. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2301.05819

[22] Gong, L. Y., & Li, X. J. (2024). A Contemporary Survey on Deepfake Detection: Datasets, Algorithms, and Challenges. Electronics, 13(3), 585-. https://doi.org/10.3390/electronics13030585

[23] Triaridis, K., & Mezaris, V. (2023). Exploring Multi-Modal Fusion for Image Manipulation Detection and Localization. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2312.01790

[24] Zampoglou, M., Papadopoulos, S., & Kompatsiaris, Y. (2017). Large-scale evaluation of splicing localization algorithms for web images. Multimedia Tools and Applications, 76(4), 4801–4834. https://doi.org/10.1007/s11042-016-3795-2

[25] Farid, H. (2009). Exposing Digital Forgeries From JPEG Ghosts. IEEE Transactions on Information Forensics and Security, 4(1), 154–160. https://doi.org/10.1109/TIFS.2008.2012215

[26] Farid, H. (2008). Digital image forensics. Scientific American, 298(6), 66–71. https://doi.org/10.1038/scientificamerican0608-66

[27] Popescu, A. C., & Farid, H. (2005). Exposing digital forgeries by detecting traces of resampling. IEEE Transactions on Signal Processing, 53(2), 758–767. https://doi.org/10.1109/TSP.2004.839932

[28] Farid, H. (2016). Photo Forensics. The MIT Press. https://doi.org/10.7551/mitpress/10451.001.0001

[29] Mahdian, B., & Saic, S. (2009). Using noise inconsistencies for blind image forensics. Image and Vision Computing, 27(10), 1497–1503. https://doi.org/10.1016/j.imavis.2009.02.001

[30] Black Lives Matter. (n.d.). Snopes. https://www.snopes.com/fact-check/blacklivesmatter/

[31] Shouldering the burden. (n.d.). Snopes. https://www.snopes.com/fact-check/shouldering-the-burden/#photo01

[32] Zampoglou, M., Papadopoulos, S., & Kompatsiaris, Y. (2017). Large-scale evaluation of splicing localization algorithms for web images. Multimedia Tools and Applications, 76(4), 4801–4834. https://doi.org/10.1007/s11042-016-3795-2

[33] Fontani M, Bianchi T, de Rosa A, Piva A, Barni M (2013) A framework for decision fusion in image forensics based on Dempster–Shafer theory of evidence. IEEE Transactions on Information Forensics and Security 8:593–607

[34] SciPy. (n.d.). scipy.stats.ks_2samp. Retrieved from https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ks_2samp.html#scipy.stats.ks_2samp

[35] Nahm, F. S. (2022). Receiver operating characteristic curve: overview and practical use for clinicians. Korean Journal of Anesthesiology, 75(1), 25–36. https://doi.org/10.4097/kja.21209

[36] B. Wen, Y. Zhu, R. Subramanian, T. Ng, X. Shen, and S. Winkler, "COVERAGE - A Novel Database for Copy-Move Forgery Detection," in Proc. IEEE Int. Conf. Image Processing (ICIP), 2016.

[37] Novozamsky, A., Mahdian, B., & Saic, S. (2020). IMD2020: A Large-Scale Annotated Dataset Tailored for Detecting Manipulated Images. 2020 IEEE Winter Applications of Computer Vision Workshops (WACVW), 71–80. https://doi.org/10.1109/WACVW50321.2020.9096940

[38] Heller, S., Rossetto, L., & Schuldt, H. (2018). The PS-Battles Dataset - an Image Collection for Image Manipulation Detection. arXiv.Org. https://doi.org/10.48550/arxiv.1804.04866

[39] Huh, M., Liu, A., Owens, A., & Efros, A. A. (2018). Fighting Fake News: Image Splice Detection via Learned Self-Consistency. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1805.04096

[40] Hsu, Y.-F., & Chang, S.-F. (2006). Detecting Image Splicing using Geometry Invariants and Camera Characteristics Consistency. 2006 IEEE International Conference on Multimedia and Expo, 549–552. https://doi.org/10.1109/ICME.2006.262447

[41] Guillaro, F., Cozzolino, D., Sud, A., Dufour, N., & Verdoliva, L. (2023). TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2212.10957

[42] IEEE IFS-TC Image Forensics Challenge. (2013). 1st Image Forensics Challenge - IEEE IFS-TC Image Forensics Challenge. Retrieved from https://web.archive.org/web/20171013200331/http://ifc.recod.ic.unicamp.br/fc.website/index.py?sec=5

[43] The Onion. (n.d.). The Onion. Retrieved from https://www.theonion.com/

[44] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In European Conference on Computer Vision (ECCV) (pp. 740–755).

[45] Wu, H., Zhou, J., Tian, J., & Liu, J. (2022). Robust image forgery detection over online social network shared images. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[46] Aniano Porcile, G. J., Gindi, J., Mundra, S., Verbus, J. R., & Farid, H. (2024). Finding AI-Generated Faces in the Wild. arXiv. https://doi.org/10.48550/arxiv.2311.08577

[47] Corvi, R., Cozzolino, D., Zingarini, G., Poggi, G., Nagano, K., & Verdoliva, L. (2023). On The Detection of Synthetic Images Generated by Diffusion Models. ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1–5. https://doi.org/10.1109/ICASSP49357.2023.10095167

[48] Mundra, S., Aniano Porcile, G. J., Marvaniya, S., Verbus, J. R., & Farid, H. (2023). Exposing GAN-Generated Profile Photos from Compact Embeddings. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 884–892. https://doi.org/10.1109/CVPRW59228.2023.00095

[49] Li, Y., Chang, M. C., & Lyu, S. (2018). In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking. IEEE International Workshop on Information Forensics and Security (WIFS).

[50] Do, N.-T., Na, I.-S., & Kim, S.-H. (2018). Forensics face detection from GANs using convolutional neural network. In ISITC..

[51] Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. In 2017 IEEE International Conference on Computer Vision (ICCV) (pp. 843–852). https://doi.org/10.1109/ICCV.2017.97

[52] Hestness, J., Narang, S., Ardalani, N., Diamos, G., Heewoo Jun, Kianinejad, H., Md Mostofa Ali Patwary, Yang, Y., & Zhou, Y. (2017). Deep learning scaling is predictable, empirically. arXiv.Org. https://doi.org/10.48550/arXiv.1712.00409

[53] Bosch, J., Olsson, H. H., Brinne, B., & Crnkovic, I. (2022). AI Engineering: Realizing the Potential of AI. IEEE Software, 39(6), 23–27. https://doi.org/10.1109/MS.2022.3199621

[54] Mareen, H., De Neve, L., Lambert, P., & Van Wallendael, G. (2024). Harmonizing Image Forgery Detection & Localization: Fusion of Complementary Approaches. Journal of Imaging, 10(1), 4-. https://doi.org/10.3390/jimaging10010004

[55] Suratkar, S., Bhiungade, S., Pitale, J., Soni, K., Badgujar, T., & Kazi, F. (2023). Deep-fake video detection approaches using convolutional – recurrent neural networks. Journal of Control and Decision, 10(2), 198–214. https://doi.org/10.1080/23307706.2022.2033644

[56] Farid, H. (n.d.). The dangers of algorithmic justice. TED Talks. Retrieved from https://www.ted.com/talks/hany_farid_the_dangers_of_algorithmic_justice?subtitle=en

[57] Amped Software. (n.d.). Amped Software. Retrieved from https://ampedsoftware.com/

[58] Cognitech. (n.d.). Cognitech. Retrieved from https://cognitech.com/

[59] FotoForensics. (n.d.). FotoForensics. Retrieved from https://fotoforensics.com/

[60] Forensically. (n.d.). Photo Forensics - Forensic Magnifier. Retrieved from https://29a.ch/photo-forensics/#forensic-magnifier

[61] Ghiro. (n.d.). Ghiro. Retrieved from https://getghiro.org/

[62] MEVER. (n.d.). Forensics. Retrieved from https://mever.iti.gr/forensics/about.html

[63] Bartoli, G. (n.d.). Sherloq [GitHub repository]. Retrieved from https://github.com/GuidoBartoli/sherloq/

[64] Krawetz, N. (2008). A Picture's Worth: Digital Image Analysis and Forensics. https://blackhat.com/presentations/bh-dc-08/Krawetz/Whitepaper/bh-dc-08-krawetz-WP.pdf

[65] Smith, A. (2008). NIGERIAN SCAM E-MAILS AND THE CHARMS OF CAPITAL. Cultural Studies, 23(1), 27–47. https://doi-org.kuleuven.e-bronnen.be/10.1080/09502380802016162

[66] Pan, S., Cui, J., & Mou, Y. (2023). Desirable or Distasteful? Exploring Uncertainty in Human-Chatbot Relationships. International Journal of Human–Computer Interaction, 1–11. https://doi-org.kuleuven.e-bronnen.be/10.1080/10447318.2023.2256554

[67] Gazzah, S., Haddada, L. R., Shallal, I., & Amara, N. E. B. (2023). Digital Image Forgery Detection with Focus on a Copy-Move Forgery Detection: A Survey. 2023 International Conference on Cyberworlds (CW), 240–247. https://doi.org/10.1109/CW58918.2023.00042

[68] Farid, H. (2019). Fake photos. The MIT Press.

[69] Stable Diffusion. (n.d.). Stable Diffusion Online. Stablediffusionweb.com. https://stablediffusionweb.com/

[70] DALL-E 3. (n.d.) Open AI. openai.com. https://openai.com/index/dall-e-3/

[71] Midjourney. (2024). Midjourney. Midjourney. https://www.midjourney.com/home

[72] Dong, C., Kumar, A., & Liu, E. (2022). Think Twice Before Detecting GAN-generated Fake Images from their Spectral Domain Imprints. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 7855–7864. https://doi.org/10.1109/CVPR52688.2022.00771

[73]van der Velden, B.H.M. Explainable AI: current status and future potential. Eur Radiol 34, 1187–1189 (2024). https://doi.org/10.1007/s00330-023-10121-4

[74] Saeed, W., & Omlin, C. (2023). Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. Knowledge-Based Systems, 263, 110273-. https://doi.org/10.1016/j.knosys.2023.110273

[75]Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1, 206–215 (2019). https://doi.org/10.1038/s42256-019-0048-x

[76] Mamun, M., Al-Kadi, M., & Marufuzzaman, M. (2013). Effectiveness of Wavelet Denoising on Electroencephalogram Signals. Journal of Applied Research and Technology, 11(1), 156–160. https://doi.org/10.1016/S1665-6423(13)71524-4

[77] Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Series B, Methodological, 39(1), 1–38. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x