Innovative Approaches to Deepfake Image Attribution: Reverse Engineering Generative Models for Enhanced Accuracy

October 2024

Authors

Parth Bhalodiya

Dr Scott Linfoot

Prepared by:

Tekh Limited

2 Newlands Road

Haconby

Bourne

PE10 0UT

United Kingdom

T: +44 (0)773 223 8192

E: info@tekh.co.uk

W: www.tekh.co.uk

Contents

Executive Summary

In the rapidly evolving landscape of digital media, the challenge of deepfake detection has garnered significant attention from major companies. However, a critical aspect often overlooked is image attribution, which is essential for law enforcement, defence, and intelligence agencies. This whitepaper addresses the pressing need for effective attribution methods, highlighting the challenges and potential solutions in this area. We propose a comprehensive approach that integrates advanced techniques to enhance accountability and traceability of manipulated content. We invite the opportunity to discuss these findings further, as we believe that collaboration is key to addressing this critical issue.

1. Introduction

In an age where seeing is no longer believing, the rise of AI-generated media has fundamentally altered the landscape of digital content. While deepfakes can manipulate reality with uncanny precision, they are not inherently negative. In fact, these synthetic creations offer remarkable opportunities for artistic expression, entertainment, gaming, and education. Rather than restricting their use, we should embrace deepfakes as innovative tools that can enhance creativity and visualise ideas in unprecedented ways.

However, as the line between real and fake blurs, society faces a pressing dilemma: how do we harness this creative potential without falling prey to the darker uses of deepfakes? Deepfakes represent a profound societal threat, as they can erode public trust in media, discredit information sources, and incite violence or panic through the spread of disinformation. As the capability to generate these images becomes more widespread, so does the potential for their malicious misuse.

To effectively navigate this dual-edged sword, we must implement robust detection methods alongside advanced attribution techniques. This dual approach not only identifies digital fabrications but also traces them back to their origins, ensuring accountability for misuse. By fostering collaboration between technology developers, policymakers, and law enforcement, we can create a framework that promotes responsible use of deepfake technology.

According to a report by DeepMedia[1], approximately 500,000 video and voice deepfakes were shared on social media globally in 2023. This represents a significant increase from previous years, highlighting the growing prevalence of AI generated content.

1.1. Problem Statement

The rise of generative image technologies, such as Automatic1111, ComfyUI, and Fooocus, combined with easy to acquire high quality generative models such as SDXL, SD3 and Flux, has revolutionised content creation, enabling users to produce highly realistic images with minimal effort. While these advancements democratise creative expression and open new possibilities in art and design, they also present significant challenges, particularly in deepfake detection and image attribution.

Figure 1. A Flux generated image with the caption: "The landscape of a park with people mulling around on a sunny day. There is a bill board in the background advertising a soft drink with the logo "Thirsty? Drink it!" on the board in colourful letters. In the foreground is a picnic table with a variety of picnic foods. The main focus is a wine bottle half full of red wine. The bottle has a white wine lable that says "Chateau de Poisson" in ornate lettering". This took around 4 minutes to create.

Deepfakes manipulated images or videos depicting events, actions, or statements that never occurred are becoming increasingly sophisticated, especially considering tools such as ReActor allow simple, yet effective face swaps. It is crucial to recognise that deepfake technology itself is not inherently harmful. When used positively, for artistic expression, education, or entertainment, deepfakes can be powerful tools for creativity and innovation. However, when wielded with malicious intent, they can have severe consequences, including election interference, spreading misinformation, defamation, and child sexual abuse. Such misuse can undermine democratic processes, damage reputations, and pose significant threats to individuals and society.

As deepfake technology evolves, the ability to generate, refine, and perfect these manipulations has outpaced current detection methods. Traditional techniques that rely on visual cues or basic forensic analysis, are no longer sufficient. Human inspection often fails to distinguish between real and fake images, highlighting the need not just to detect alterations but to determine precisely how they were made. Identifying the specific generative model, parameters used, and workflow followed in creating a deepfake is critical for attributing them to their creators, especially when the intent is to deceive or cause harm.

In 2020, researchers found over 85,000 harmful deepfake videos online, a number that has been doubling every six months[1]. By 2023, the number of deepfake videos online had increased by 550% compared to 2019[1].

[1] https://contentdetector.ai/articles/deepfake-statistics/

1.2. The Critical Imperative: Advanced Deepfake Attribution for National Security

Detecting deepfakes is only the first step; tracing the origin of manipulated content is equally important. To effectively combat misinformation and disinformation and hold threat actors accountable, image attribution identifying the source and specific tools used in creating a deepfake is essential. Reverse engineering the generative process plays a pivotal role in unmasking the origins of deepfakes and preventing their harmful misuse.

Why Are We Doing This? The stakes in deepfake detection and attribution could not be higher, especially for organisations where the authenticity of visual information is paramount. For law enforcement agencies, including defence departments and national security organisations, the ability to distinguish between genuine and fake contents are no longer a luxury, it is a necessity. The consequences of failing to detect and attribute deepfakes could be catastrophic, leading to compromised public safety and national security. Deepfakes can fabricate evidence, create false narratives, or impersonate individuals in ways that have far-reaching and potentially devastating effects.

Failing to prioritise this aspect can lead to significant national security threats, as deepfakes being weaponised to create false intelligence, impersonate government officials, or spread disinformation that destabilises governments. This presents a critical concern for defence agencies tasked with safeguarding national security.

Moreover, as deepfakes become more prevalent, there is a risk of eroding public trust in digital media. If all digital content is viewed with suspicion, the credibility of genuine information is undermined, complicating efforts to communicate valuable insights. This erosion of trust can have far-reaching implications for society, making it essential to establish robust methods for attribution.

The increasing volume of deepfake content challenges existing detection systems and strains resources across various sectors, including law enforcement and media organisations. These entities must invest significant time and resources into combating the spread of deepfakes, highlighting the urgent need for effective attribution methods.

Additionally, deepfakes pose challenges in counterterrorism and intelligence operations. They can be exploited by terrorist organisations or hostile state actors to create false propaganda, spread disinformation, or disrupt intelligence efforts. Accurate attribution and detection are vital in countering these threats, ensuring that we can identify the sources of malicious content and hold them accountable.

By focusing on both detection and attribution, we can mitigate these risks and foster a safer digital environment, ultimately preserving the integrity of information and protecting democratic processes

1.1.1. Ensuring Legal and Investigative Integrity

In legal contexts, where digital evidence is increasingly scrutinised, the ability to attribute an image to its source and understand how it was created is indispensable. Without this capability, it becomes difficult to prove the authenticity or inauthenticity of digital content, which could lead to wrongful convictions, legal disputes, or the undermining of justice. Robust attribution and reverse engineering provide the necessary evidence to support legal and investigative processes.

1.1.2. Election Integrity

Imagine a deepfake video showing a political leader making inflammatory statements or confessing to a crime they never committed. In the heat of an election, such a fabrication could shift voter opinion, disrupt democratic processes, and potentially lead to the election of unqualified or malevolent candidates. This is not a hypothetical scenario it is a very real threat that could undermine the foundations of democracy. All one needs to do is to look at the Trump/Harris elections in the US and the “Swifties for Trump” incident[2].

1.1.3. CSAM

Beyond elections, the implications of deepfakes extend to areas as grave as the proliferation of Child Sexual Abuse Material (CSAM). The potential for deepfake technology to create fabricated pornographic content involving minors exacerbates an already critical issue, presenting new and horrifying threats to innocent individuals whose likenesses are exploited. One incident reported[3] that a child was rescued from her abuser, but now new images of this child are being created by individuals who want to see her in new, abusive, situations, years after the physical abuse ended. For this child, the abuse has not ended and will be reminded of this throughout her whole life.

So What? If we fail to develop more advanced methods for deepfake detection and, crucially, for image attribution, the consequences could be dire. Detection alone is insufficient it merely tells us that something is wrong. What is needed is attribution, the ability to trace a deepfake back to its source. This capability is vital for law enforcement and intelligence agencies as it allows them to identify the perpetrators behind these manipulations and understand their motives.

1.2. The Importance of Attribution Over Detection

In the battle against deepfakes, many players, including tech giants like Google and Meta, have invested heavily in detection technologies. However, attribution pinpointing the exact origins of these manipulated images or videos remains an unsolved and critically important problem. For intelligence and law enforcement agencies, attribution is the key to moving from merely understanding what has been altered to identifying who is responsible for the

[2] https://www.cbsnews.com/news/trump-shares-fake-swifties-for-trump-images/
[3] https://www.iwf.org.uk/media/nadlcb1z/iwf-ai-csam-report_update-public-jul24v13.pdf

manipulation. Without accurate attribution, it is nearly impossible to trace an attack back to its source, understand the intent behind it, or take preventive actions against future threats.

Attribution goes further by:

Identifying Threat Actors: Pinpointing the exact source of a deepfake can reveal whether the manipulation was the work of state actors, organised crime, hacktivists, or just malicious people (the “script kiddies”), which is crucial for developing appropriate responses.

Mitigating Future Attacks: Understanding the methods and tools used in the creation of deepfakes allows for the development of more targeted and effective countermeasures. This knowledge is invaluable in disrupting the workflow of malicious actors.

Strengthening National Security: Attribution provides insights into the larger geopolitical or criminal context of a disinformation attack, offering information critical to national security operations.

Why Is This an Unsolved Problem? Despite the importance of attribution, the current focus in the market has been disproportionately on detection. This leaves a significant gap in our ability to address the more strategic questions of where deepfakes come from and who is responsible. Tackling this challenge requires not just technological innovation but a paradigm shift towards forensic analysis techniques that can unravel the origins of deepfakes.

1.3. The Key to Solving the Problem

A shift in focus towards developing robust image attribution capabilities is essential for several critical reasons:

1. **Identification of Source and Accountability:** By identifying the specific generative model, parameters, and workflow used to create a deepfake, law enforcement and security agencies can trace the manipulation back to its source.
2. **Understanding and Mitigating Malicious Intent:** Whether the deepfake was created for political manipulation, disinformation campaigns, or personal vendettas, understanding the source helps in addressing the root cause of the threat, not just its symptoms.
3. **Keeping Pace with Technological Evolution:** Deepfake technology is evolving rapidly, with new models and techniques being developed continuously. Traditional detection methods, which may rely on visual cues or simple forensic analysis, struggle to keep up with these advancements. Understanding these new techniques enables the development of more sophisticated and adaptive detection strategies that can keep pace with the evolution of generative models.
4. **Policy:** To effectively combat the challenges posed by deepfakes and misinformation, it is essential to establish clear policies that draw valuable lessons from the cybersecurity sector. Just as cyber threats are managed through comprehensive frameworks, similar strategies must be applied to digital deception. Developing policies that define ethical standards for content creation and requiring transparency in the use of deepfake technology can help mitigate risks. Regulatory

measures should be implemented to ensure that creators and platforms are held accountable for the content they produce, fostering a culture of responsibility in the digital landscape.

5. **Provenance**: Robust provenance tracking is crucial for managing the risks associated with deepfakes. Techniques such as digital watermarking can help identify the source of content, while blockchain technology can provide an immutable record of the creation and modification of digital media. This enhances accountability and traceability, ensuring that any alterations to content are documented and verifiable. By treating misinformation and disinformation as serious threats akin to cyber threats, we can adopt threat assessment frameworks to evaluate their potential impact on public trust and national security. Establishing incident response plans for misinformation crises will ensure that organisations are prepared to act swiftly when deepfakes are identified, minimising their societal impact.

Conclusion: The Urgent Need for a Paradigm Shift

While most current efforts are focused on deepfake detection, it is imperative to extend our capabilities to include robust image attribution and reverse engineering. This shift is not just an option; it is a necessity in the face of the growing challenges posed by deepfakes. These advanced capabilities are crucial for ensuring accountability, understanding malicious intent, adapting to technological advancements, supporting legal processes, counteracting future threats, and maintaining public trust and national security.

2. Background Information: Context and Related Work

2.1. The Evolution of Generative Image Technologies

The landscape of generative image technologies has undergone rapid and transformative changes in recent years. Initially, the ability to create realistic images through artificial intelligence was confined to specialised research labs and tech giants. However, the advent of user-friendly tools like Automatic1111, ComfyUI, and Fooocus[4] has democratised this technology, allowing virtually anyone with a computer to produce high-quality, photorealistic images with minimal effort[5]. These advancements have spurred innovation across various fields, including entertainment, advertising, and art, by offering unprecedented creative freedom. However, the same ease of use those fuels creativity also lowers the barrier for misuse, leading to the proliferation of deepfakes[6].

2.2. The Emergence and Impact of Deepfakes

Deepfakes, which first gained widespread attention in 2017, have quickly evolved from a novelty into a significant societal concern[7]. Powered by deep learning techniques, particularly Generative Adversarial Networks (GANs), deepfakes enable the creation of hyper-realistic videos and images that can convincingly mimic real people, often placing them in situations or actions that never occurred. While the technology itself is neutral, its applications can be either benign or malicious. On the positive side, deepfakes have been utilised for creative projects[8], such as resurrecting historical figures in documentaries or enhancing cinematic effects[9]. However, the darker side of deepfakes has drawn considerable scrutiny due to their potential for misinformation, defamation, and the creation of non-consensual explicit content, all of which carry potentially devastating consequences for individuals and society[10].

2.3. Challenges in Deepfake Attribution

Attributing deepfakes to their source models presents unique challenges. Unlike detection, which focuses on identifying whether an image or video is fake, attribution aims to trace the origin of the fake content back to the specific generative model used[11]. This process involves identifying unique fingerprints or artifacts left by different models, which can be subtle and difficult to detect. The rapid evolution of generative models further complicates

[4] https://www.digitalocean.com/community/tutorials/achieving-the-highest-fidelity-image-synthesis-with-fooooocus
[5] https://game.intel.com/us/stories/comfyui-vs-fooocus-for-genai-on-intel-arc-gpus/
[6] https://www.technologyreview.com/2023/10/04/1080801/generative-ai-boosting-disinformation-and-propaganda-freedom-house/
[7] https://www.academia.edu/74762342/The_Emergence_of_Deepfakes_and_its_Societal_Implications_A_Systematic_Review
[8] https://theconversation.com/deepfakes-are-being-used-for-good-heres-how-193170
[9] https://www.analyticsinsight.net/deepfake-technology/deepfake-in-entertainment-impact-on-film-and-television
[10] https://journals.sagepub.com/doi/pdf/10.1177/2056305120903408
[11] https://www.computer.org/csdl/journal/tp/2023/12/10202583/1PhSMobmlUs

attribution, as new models continuously emerge with improved capabilities and different characteristics[12].

Current attribution methods rely on advanced machine learning techniques to analyse these fingerprints[13]. However, these methods face limitations in terms of scalability and robustness. The diversity of generative models and the variety of ways they can be used to create deepfakes mean that attribution algorithms must be highly adaptable and capable of generalising across different scenarios. Additionally, the lack of standardised datasets and evaluation metrics for attribution poses a significant challenge, making it difficult to compare the effectiveness of different approaches[14].

2.4. Reverse Engineering of Deceptions (RED)

A significant advancement in the field of image attribution is the concept of Reverse Engineering of Deceptions (RED), as explored by Yao et al. (2024)[15]. RED represents a novel and dynamic approach within adversarial machine learning, focusing on both machine-centric and human-centric attacks. For machine-centric attacks, RED involves reverse engineering methods to uncover pixel-level perturbations, adversarial saliency maps, and victim model information embedded within adversarial examples. This allows for a deeper understanding of the adversary's knowledge, objectives, and the specifics of their attack toolchains.

In the realm of human-centric attacks, RED focuses on generative model information inference and manipulation localisation from generated images. Techniques such as model parsing extracting hyperparameters used in generative models and manipulation localisation identifying tampered regions within images are pivotal in attributing deepfakes to their sources. These methodologies not only enhance the ability to detect deepfakes but also provide insights into the adversarial strategies employed, thereby strengthening the overall defence mechanisms against digital deceptions.

2.5. DE-FAKE: Detection and Attribution of Fake Images Generated by Text-to-Image Generation Models

The DE-FAKE study by Sha et al. (2023)[16] addresses the detection and attribution of fake images generated by text-to-image models like DALL·E 2, Stable Diffusion, GLIDE, and Latent Diffusion. The researchers developed a machine learning classifier capable of distinguishing fake images from real ones by identifying common artifacts shared by different models. Furthermore, they demonstrated that fake images could be effectively attributed to their source models, as each model leaves unique fingerprints in the images it generates. This attribution capability is crucial for holding model owners accountable for the misuse of their technologies.

[12] https://arxiv.org/html/2407.06174v4
[13] https://github.com/vishal3477/Reverse_Engineering_GMs
[14] https://www.mdpi.com/2073-431X/12/10/216
[15] http://dx.doi.org/10.1561/3300000039
[16] https://dl.acm.org/doi/abs/10.1145/3576915.3616588?casa_token=kJrGjjHNw14AAAAA:QmgL8eMPvaaUjfJfzTfifK9T3FK_KJEaQHguo8VxY9tcR2CadsZl-BzwlJENkXTZvEdNrdptqg

2.6. Rethinking Open-World DeepFake Attribution with Multi-Perspective Sensory Learning (MPSL)

Sun et al. (2024)[17] introduced the concept of Open-World DeepFake Attribution (OW-DFA++) and proposed the Multi-Perspective Sensory Learning (MPSL) framework to address the challenges of attributing deepfakes in open-world scenarios. The MPSL framework employs a Multi-Perception Voting (MPV) module to align inter-sample features based on global, multi-scale local, and frequency relations. This approach effectively groups samples belonging to the same attack type, enhancing the accuracy of deepfake attribution. Additionally, the Confidence Adaptive Pseudo labelling (CAP) module mitigates pseudo-noise and improves class compactness by adaptively filtering samples with high uncertainty.

2.7. The Path Forward

As deepfake technology continues to advance, the gap between detection capabilities and the sophistication of generated content is likely to widen. Addressing this challenge requires a multi-faceted approach that combines detection with robust attribution techniques. By focusing on the reverse engineering of generative models and enhancing forensic capabilities through frameworks like RED, we can improve our ability to not only detect deepfakes but also to hold their creators accountable. This shift is crucial for maintaining the integrity of digital content in an era where the line between reality and fabrication is increasingly blurred.

Future research should prioritise the integration of machine-centric and human-centric attribution methods, leveraging advancements in adversarial machine learning and trustworthy computer vision. The DE-FAKE study by Sha et al. (2023) has demonstrated the potential of using unique fingerprints left by different generative models to attribute fake images to their source models. Building on this, further refinement and application of these techniques can enhance our ability to trace the origins of deepfakes accurately.

The Multi-Perspective Sensory Learning (MPSL) framework introduced by Sun et al. (2024) offers a promising approach to deepfake attribution in open-world scenarios. By aligning inter-sample features based on global, multi-scale local, and frequency relations, MPSL improves the accuracy of deepfake attribution. Future research should explore the integration of MPSL with other attribution methods to create more robust and scalable solutions.

Collaborative efforts between industry and governmental bodies will be essential to develop scalable and effective solutions. Additionally, exploring RED could offer innovative pathways for immutable and traceable deepfake attribution, further strengthening our defences against digital deceptions. By addressing these multifaceted challenges through comprehensive and forward-looking strategies, we can better safeguard the authenticity of digital media, protect individual reputations, and uphold the foundational pillars of democratic processes and societal trust. Integrating insights from studies like DE-FAKE and MPSL will be instrumental in advancing the field and mitigating the risks associated with deepfakes.

[17] https://link.springer.com/article/10.1007/s11263-024-02184-7

3. Solution overview

This project introduces an innovative approach to content attribution in digital media. By leveraging advanced techniques in feature extraction and reverse engineering, this solution addresses the increasing need for reliable identification of manipulated content and its origins. Our methodology integrates sophisticated fingerprint analysis with machine learning and genetic algorithms to create a robust attribution system.

While we are still in the early stages of our research, our system is designed to explore the potential of tracing manipulated content back to its generative source. This process aims to enhance our understanding of the technologies involved in digital creations, providing a foundation for future developments in attribution methods.

3.1. Technology and Methodology

3.1.1. Advanced Fingerprint Estimation and Feature Extraction

Every image or video created by a generative model, carries subtle but detectable traces known as “fingerprints.” These fingerprints are unique to each generative model and can be compared to the digital equivalent of human fingerprints. They serve as critical indicators that help us trace back to the model responsible for creating the content.

As we are in the early stages of our research, our system is designed to analyse these generative fingerprints using a variety of advanced techniques. This includes not only spatial and frequency domain analysis but also other innovative methods.

By combining these diverse methodologies, we aim to facilitate comprehensive attribution. This multifaceted approach allows us to explore the unique characteristics of each image or video, enhancing our understanding of the generative processes involved and improving our ability to trace manipulated content back to its source.

A. Spatial Domain Techniques

In this approach, our system examines pixel-level patterns that vary between real and fake images. The **ResNet50 model**, pre-trained on ImageNet, is employed to extract deep latent features from the input images. By removing the last fully connected layer of the ResNet50, we isolate high-level features that capture the texture, structure, and overall pixel distribution in the image.

The system looks for inconsistencies in pixel arrangement that can reveal signs of tampering, such as **blended edges**, **unnatural lighting transitions**, or other artifacts introduced during manipulation. The model processes the image using a **feature extractor**, which flattens and standardises the latent features for subsequent analysis. By extracting and analysing these deep spatial features, our system ensures high accuracy in detecting subtle changes that indicate manipulation.

B. Frequency Domain Techniques

In addition to spatial domain analysis, our system applies **frequency domain techniques** to detect hidden anomalies. Using transformations like the **Discrete Cosine Transform (DCT)**, the image is converted from the spatial domain into the frequency domain, where periodic patterns become more visible. The DCT helps us spot irregularities that are not apparent through pixel-level analysis, such as **hidden noise** or **distortions** introduced by generative models.

Furthermore, the system employs **high-pass filtering** to detect edge-based anomalies, which are common in manipulated images. The high-pass filter emphasises the image's edges and removes low-frequency components, making it easier to highlight any **localised artefacts** caused by deepfake creation. By comparing these frequency patterns to a library of known fingerprints from generative models, our system can accurately identify the model that likely produced the fake content.

C. Latent Feature Analysis

In addition to various analytical techniques, our system employs latent feature analysis to uncover hidden patterns within images that may indicate the generative model used. By extracting latent features, we can gain insights into the underlying structure of the content, which is crucial for effective attribution.

Using advanced machine learning models, we process the images to identify these latent features, which encapsulate complex characteristics that are not immediately visible. This analysis allows us to discern subtle differences between genuine and manipulated content, enhancing our ability to trace back to the specific generative model.

Moreover, the system integrates multiple analytical lenses, including statistical methods and deep learning approaches, to enrich the feature extraction process. By comparing the extracted latent features against a library of known fingerprints from various generative models, our system aims to identify the model that likely produced the manipulated content. This comprehensive approach not only improves attribution accuracy but also deepens our understanding of the generative processes involved.

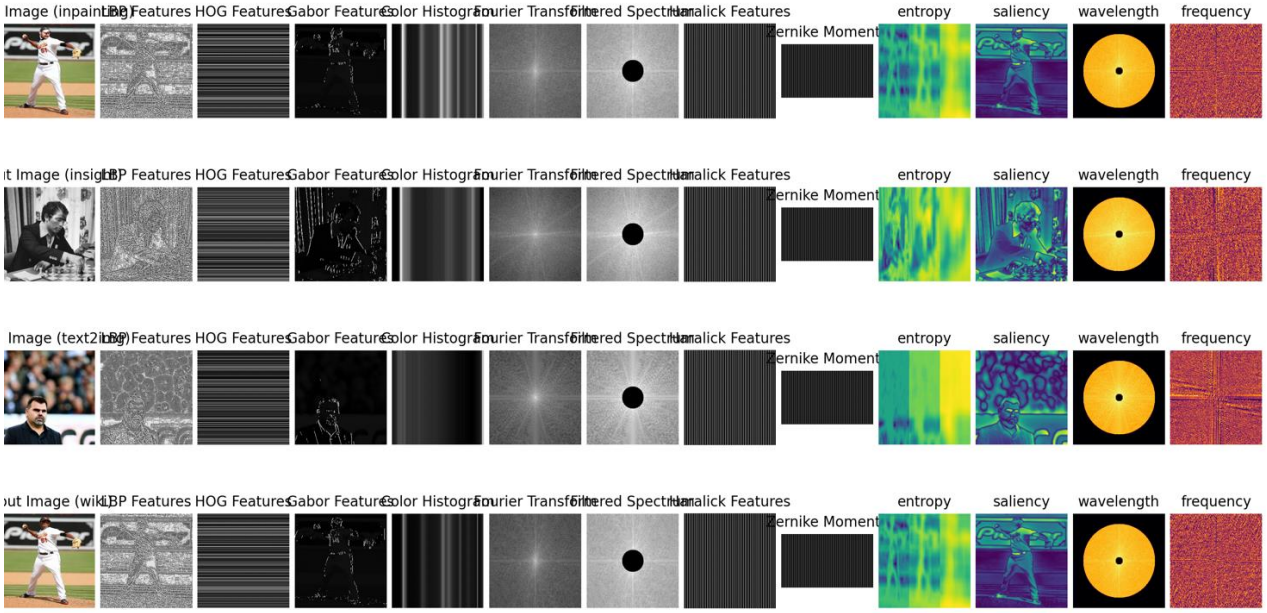

Figure 3.1. Different features for fingerprint analysis

D. Combined Feature Extraction for Robust Detection

The system's strength lies in combining spatial and frequency domain features with deep latent features. This multi-faceted approach enables us to analyse an image holistically, ensuring that both pixel-level irregularities and hidden frequency domain anomalies are considered. The extracted features from each domain are concatenated into a robust feature vector, which serves as the ba sis for deepfake detection and model attribution

This **combined feature vector** captures a rich representation of the image, making it highly effective in distinguishing between real and fake content, and attributing the generative model responsible for its creation.

3.1.2. Machine Learning Integration

Once we have extracted the combined features spatial, frequency, and other relevant characteristics we can map them onto a high-dimensional space. This mapping allows us to leverage machine learning techniques, particularly nearest neighbour classifiers, to analyse the data effectively.

By employing these classifiers, we can generate hypotheses regarding the origin of the content. For instance, we might conclude that a particular image was generated by a Stable Diffusion 3 model, or we could determine that it is a genuine image. Additionally, we can identify other generative models, such as Flux1 or StarGAN, based on the patterns observed in the feature space.

This approach not only enhances our ability to attribute manipulated content to its source but also provides a framework for understanding the nuances of different generative models. By continuously refining our classifiers and expanding our feature set, we aim to improve the accuracy and reliability of our attributions.

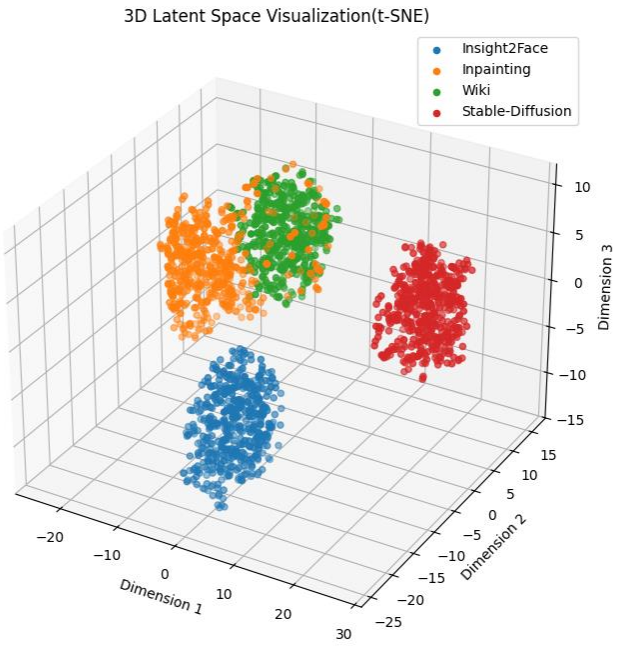

Figure 3.2. 3D Latent Space Visualization of Model Fingerprint Features (t-SNE). This plot visualizes fingerprint features from four different models: Insight2Face, Inpainting, Wiki, and Stable-Diffusion. The t-SNE technique clusters these fingerprint features, illustrating the distinct characteristics or "fingerprints" left by each model in the latent space.

3.1.3. Genetic Algorithms

To enhance attribution accuracy, the solution incorporates genetic algorithms (GAs)[18] that intelligently evolve potential model and parameter combinations. This innovative approach allows GAs to iteratively refine the search for the most likely source of a manipulated image, significantly improving the precision of the attribution process. By leveraging the adaptive nature of GAs, the system effectively identifies the origins of deepfakes, keeping pace with the rapid evolution of generative model technologies.

The methodology integrates advanced fingerprint estimation with machine learning and genetic algorithms, utilising the GA component as a brute-force method to explore a vast search space for optimal solutions. This exploration is particularly beneficial in several key areas related to attribution and reverse engineering:

A. Model Identification

Genetic Algorithms (GAs) play a crucial role in identifying the specific generative model used to create a deepfake. The process begins with the initialisation of a diverse population of potential models, each representing different configurations of the generative model. Each model undergoes a fitness evaluation based on its ability to accurately attribute manipulated images to their sources. The best-performing models

[18] https://www.sciencedirect.com/topics/engineering/genetic-algorithm

are selected to create the next generation, ensuring that only the most promising candidates are carried forward. Through crossover and mutation processes, new models are generated, introducing variability that helps the algorithm escape local optima. This iterative refinement continues over multiple generations, allowing the GA to converge on the most effective model for attribution, thereby significantly enhancing reliability.

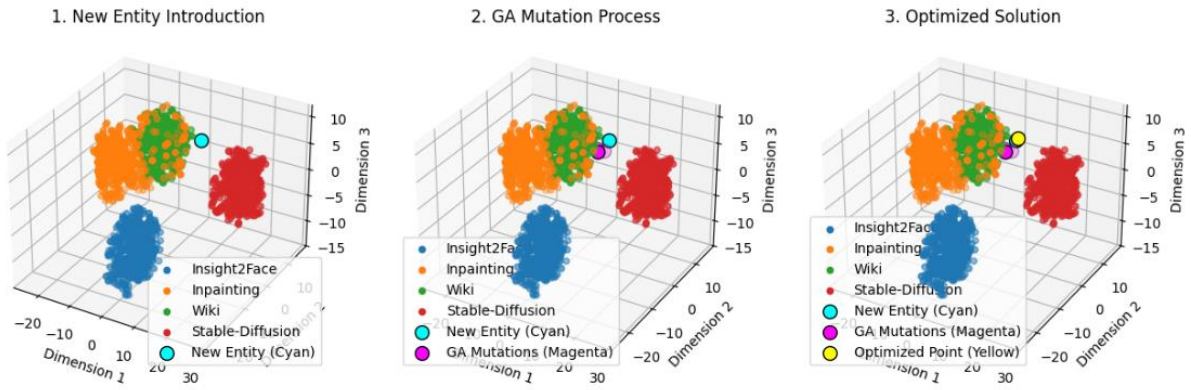

Figure 3.3. ***3D Visualization of Genetic Algorithm Optimization in Latent Space (t-SNE).*** *This series of plots shows the process of introducing a new entity and optimizing it using a genetic algorithm (GA) in the latent space of four models: Insight2Face, Inpainting, Wiki, and Stable-Diffusion.*

1. ***New Entity Introduction****: A new entity (cyan) is introduced into the latent space.*
2. ***GA Mutation Process****: Genetic algorithm mutations (magenta) are applied to explore the space and evolve the entity toward better solutions.*
3. ***Optimized Solution****: The optimized point (yellow) represents the final solution, refined through GA, achieving the best result near the clusters.*

B. Feature Selection

Feature selection is vital for enhancing the robustness and efficiency of the attribution process, and genetic algorithms provide a powerful mechanism for this. GAs evaluates the importance of different features in the dataset by assessing their contribution to the model's ability to trace back to the generative source. By identifying and retaining only the most relevant features, GAs help reduce the dimensionality of the dataset, which can lead to faster processing times and improved interpretability. The iterative nature of GAs allows them to refine the feature set by testing various combinations and selecting those that yield the best performance. This focused approach not only improves the accuracy of the attribution system but also ensures that it remains resilient against noise and irrelevant data, ultimately leading to more reliable identification of the origins of deepfakes.

Conclusion

This white paper underscores the urgent need for advanced image attribution methods in the context of deepfakes, particularly for the Ministry of Defence and law enforcement agencies. As deepfake technology evolves, the potential for misuse escalates, posing significant threats to national security and public safety. Our exploration of reverse engineering generative models reveals that robust attribution capabilities are essential for identifying the origins of manipulated content and holding malicious actors accountable.

However, while we have made strides in developing robust attribution methods, there remains much to explore. The rapid evolution of generative technologies presents ongoing challenges that require continuous innovation and collaboration among stakeholders. As we refine our techniques, it is essential to address the complexities of deepfake creation and the diverse motivations behind their use.

By fostering dialogue and partnerships across technology, policy, and law enforcement sectors, we can enhance our understanding and effectiveness in combating the misuse of deepfakes. This proactive approach will not only strengthen our current frameworks but also pave the way for future advancements in ensuring accountability and preserving public trust in digital media.