Exploring Dataset Diversity for GenAI Image Inpainting Localisation in Digital Forensics

Matthew Thomson
Edinburgh Napier University
Edinburgh, Scotland, United Kingdom
matthew.thomson@napier.ac.uk

Sean McKeown
Edinburgh Napier University
Edinburgh, Scotland, United Kingdom
s.mckeown@napier.ac.uk

Richard Macfarlane
Edinburgh Napier University
Edinburgh, Scotland, United Kingdom
r.macfarlane@napier.ac.uk

Petra Leimich
Edinburgh Napier University
Edinburgh, Scotland, United Kingdom
p.leimich@napier.ac.uk

Abstract

Generative Artificial Intelligence (GenAI) has significantly increased the sophistication and ease of image tampering techniques, posing challenges for digital forensics in identifying manipulated images. A lack of dataset standardisation hinders the ability to effectively benchmark and compare GenAI inpainting localisation techniques, reducing their reliability in digital forensic applications. This paper aims to address this gap by exploring the need for standardised criteria for datasets in digital forensics for benchmarking detection techniques through preliminary experiments.

To address the limited diversity in existing datasets, a small-scale dataset was developed, consisting of 240 tampered images, 20 masks and 20 authentic images. This dataset includes four subject image classes (animals, objects, persons, scenery) and three inpainting tools (GLIDE, GalaxyAI, Photoshop). The dataset was evaluated against 13 localisation algorithms from the Image Forensics MATLAB Toolbox to determine key components that should be considered in the standardisation of testing environments.

The results show that the images in the animals and persons categories achieved the highest F1-Scores and accuracy over the other classes. Among tools, GLIDE inpainted images were consistently shown to be the most challenging to detect, underscoring the importance of further investigating these images. These findings provide foundational insights for identifying a set of criteria to establish robust testing environments, enabling the development of reliable and accurate GenAI inpainting localisation techniques.

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**; • **Applied computing** → *Computer forensics.*

Keywords

Digital Forensics, Artificial Intelligence (AI), Generative AI (GenAI), AI Manipulation, Inpainting, Image Forgery Localisation

DFDS 2025, Brno, Czech Republic

ACM ISBN 979-8-4007-1076-6/25/04
https://doi.org/10.1145/3712716.3712724

ACM Reference Format:
Matthew Thomson, Sean McKeown, Richard Macfarlane, and Petra Leimich. 2025. Exploring Dataset Diversity for GenAI Image Inpainting Localisation in Digital Forensics. In *Digital Forensics Doctoral Symposium (DFDS 2025), April 01, 2025, Brno, Czech Republic.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3712716.3712724

1 Introduction

Image tampering has become increasingly sophisticated, posing significant challenges to digital forensics, especially with the introduction of generative artificial Intelligence (GenAI). GenAI is now capable of creating media that appear highly realistic: impossible for the human eye to distinguish tampered from authentic images, and difficult to detect automatically. If undetected, such media could lead to misinformation impacting the investigation process, leading to wrongful convictions or releasing a guilty suspect. These challenges highlight the need for work towards robust and reliable tampering detection and localisation methods [17, 22].

Realistic testing environments are essential to accurately evaluate detection and localisation methods. Without diverse and realistic datasets, the accuracy and reliability of detection and localisation methods cannot be effectively evaluated. To overcome these challenges, this study aims to identify the similarities and patterns within subject image classes and tools that contribute to the performance of detection and localisation techniques through a quantitative evaluation. The experimentation considered the subject class of an image, as the variation in image complexities and textures can significantly affect performance. Additionally, the inpainting tool utilised introduces differing image attributes due to its divergent image manipulation processes. Through preliminary experimentation, this work demonstrates how these components impact localisation results, emphasising the importance of standardised and realistic testing environments.

To address the challenges, we explore the following research questions:

- **RQ1:** How do different image classes, such as animals, objects, persons, and scenery, affect the performance of tampering localisation algorithms?
- **RQ2:** How do different inpainting tools, such as GLIDE, GalaxyAI, and Photoshop, impact tampering localisation performance?

- **RQ3:** What additional considerations should be addressed to establish a comprehensive set of criteria for the standardisation of a realistic testing environment for digital forensics purposes?

Our evaluation lays the foundation towards establishing standardised criteria for testing environment datasets within digital forensics. Developing a standardised dataset will enable consistent testing, improving the reliability and applicability of GenAI image detection and localisation techniques in forensic investigations.

2 Background and Related Work

In digital forensics, image tampering or forgery refers to the intentional modification of the content to mislead the viewer. Common image tampering techniques include copy-move and splicing manipulations [24]. However, the integration of GenAI into these processes has significantly increased their sophistication, making them particularly dangerous. Alongside this, image tampering is now more accessible and requires no prior expertise, widening its potential for use. GenAI for image manipulation can typically be divided into two main categories: fully generated and partially generated content. Partially AI-generated or tampered images involve the use of GenAI models to enhance traditional manipulation techniques [23]. One widely accessible and commercialised image tampering technique is inpainting, where elements of interest, such as people or weapons, can be removed from the scene to alter the image's context [23]. Malicious image tampering such as this can be used to obscure digital evidence, posing significant challenges for forensic investigators [25]. The ability of GenAI models to generate highly realistic and coherent background textures from object removal makes detecting tampered regions particularly difficult.

While new detection and localisation techniques are being developed, their evaluation is often limited by the lack of standardised criteria for testing environments. For instance, Li et al. [12] found a substantial drop in performance when testing their transformer-based detection technique against unseen datasets. Similarly, Patel et al. [19] explored the use of machine-based techniques, specifically Dense CNNs, and evaluated their approach using the Deepfake Images Detection and Reconstruction Challenge dataset. Patel et al.'s study reported the accuracy metric ranging from 94.67% to 99.33%. However, when testing the generalisation capabilities of the proposed model, the accuracy decreased significantly to only 77%. This highlights the importance of standardised datasets and testing environments in determining a technique's applicability to digital forensics. Some techniques may perform better against certain manipulation types or GenAI models than others and this is hard to determine without a wide scale dataset for benchmarking. This lack of standardisation limits the ability to assess the accuracy and reliability of detection and localisation techniques across diverse scenarios. Forensic analysts often encounter highly specific manipulation types that correspond to different image classes, such as face swaps in social media or object removal in CCTV [16]. A testing environment that allows for divided classes, such as persons, objects, and scenery, would enable forensics analysts to evaluate whether the nature of the replacement influences localisation performance. Identifying these impacts allows examiners to tailor their analysis strategies to specific content types.

Despite the number of datasets available in the field, their relevance becomes outdated very quickly, with new models constantly being developed. Furthermore, several of the datasets tend to be for very specific problems or areas rather than providing a diverse variety of options. Datasets such as *Artifact* [20] include images from a variety of models and classes, but their exclusive focus on art styles and the absence of options to isolate and test specific classes limits their applicability for forensic scenarios. The Hierarchical Fine-grained (*HiFi-IFDL*) dataset [8] addresses additional manipulation types beyond just fully AI-generated or inpainting, but omits commonly used commercial models such as Adobe Firefly, Midjourney, DALL-E, or newer versions of Stable Diffusion. Conversely, datasets such as *CIFAKE* [5] have a large variety of classes, but are limited by very small image resolutions of 32x32 or 64x64 compared to the typical Stable Diffusion output of 512x512.

This work aims to address such gaps by exploring the influence of dataset components, such as a variety of image classes and tools, on localisation performance. By identifying foundational components that should be dictated in standardised criteria for datasets, our work contributes to developing robust and standardised testing environments for digital forensics applications.

3 Methodology

To evaluate the requirements for a standardised testing environment for GenAI localisation techniques, we assess the impact of dataset components through two primary experiments. The first experiment focuses on the influence that the image subject classes, animals, objects, persons, and scenery, can have on the localisation performance. The second experiment evaluates the impact of the inpainting tool used for tampering, where three tools, GLIDE, GalaxyAI, and Adobe Photoshop, are chosen for their distinct approaches to inpainting. The focus on localisation is due to the availability of the Image Forensics MATLAB Toolbox [28], which contains a range of localisation algorithms. This approach allows for the evaluation to prioritise the impact of dataset components, rather than specific detection or localisation techniques. To conduct these experiments, a comprehensive dataset consisting of authentic images, tampered images, and masks, must be created, alongside adaptations to the Image Forensics MATLAB Toolbox.

3.1 Dataset Creation

A small-scale dataset consisting of 20 authentic images with 20 corresponding edit localisation masks, each used to generate 12 tampered images (totalling 240), with the tampered content being created using the GenAI technique of inpainting. This tampering method alters an image by reconstructing segments to conceal elements, often deceiving viewers [23]. The inpainting process is displayed in Figure 1, where the cow is removed from the image. This technique was selected due to its availability and ease of use, alongside similarities to the traditional technique of splicing.

Figure 1: Example of inpainting using GalaxyAI on an image of a cow.

The authentic images were sourced from the Microsoft Common Objects in Context (*MS COCO*) dataset [13], which contains 90 categories of images. Four categories, animals, objects, persons and scenery, were selected to include a variety of image options and complexities. Within each subject class, five images were selected from categories of similar classes. For example, dog, cat, bear, chicken and cow were combined to form the "animals" class. The four selected image subject classes represent a variety of the 90 *MS COCO* categories, enabling the combination of multiple subcategories to provide a more diverse evaluation.

To manipulate the images with the chosen inpainting techniques, each authentic image was tampered with four times using the three tools. The first method involved Adobe Firefly, which was accessed through the desktop Adobe Photoshop application [1]. Subject masks in the images were selected using the semi-automatic Magic Selection Tool, and the prompt "remove" was used to guide the inpainting process. Photoshop's GenAI tool generates three variations per run, so the process was repeated twice to produce sufficient outputs, with the first four results selected. The second method used GLIDE, which was run locally using the code available from the glide-text2im project [18]. To allow for automation, the Jupyter notebooks contained within the project's source code were converted into a Python script. The script requires the input of the authentic image, a reference mask, and the "remove" prompt. Masks initially created from the Photoshop process were reused to ensure consistency across tools. The third method used GalaxyAI, an object editor and removal feature on Galaxy phones [21]. The masks were manually drawn over the images to replicate those used in Photoshop and GLIDE. This tool uses an erase button to remove the selected masks, no prompt is required.

For each category, five masks corresponding to the authentic images were created, resulting in twenty masks across the dataset. These masks, initially created in Adobe Photoshop, were used throughout for the tampering reference masks. The full dataset consists of 3 tools, 4 classes, and 20 tampered images each, resulting in 240 tampered images total (3x4x20).

3.2 Experiment Setup

The experiments were conducted using the Image Forensics MATLAB Toolbox [28], which was created primarily for splicing localisation. To improve upon the work presented in [28], our experimentation includes images with inpainting modification rather than splicing. The evaluation focuses on the differing performance across classes and tools rather than each algorithm's performance. Additionally, the toolbox was used to allow for multiple localisation methods to be run on the images, providing results specific to the tampered images rather than one specific method of localisation. The toolbox contains a total of 16 algorithms, of which 13 were selected for this experiment as they can be directly applied without further modifications to the code being necessary. The selected algorithms are: ADQ1 [14], ADQ2 [3], ADQ3 [2], BLK [11], CAGI [9], CFA1 [7], CFA3 [6], DCT [27], ELA [10], NADQ [4], NOI1 [15], NOI4 [26], and NOI5 [29].

To adapt the toolbox, the EvaluateAlgorithm.m script was modified to allow for multiple algorithms to be passed at once, and the ExtractMaps.m script was modified to allow for an option accommodating the specific mask and authentic image setup that this experiment consisted of. Additional MATLAB scripts were created to process the probability maps generated by the toolbox, which represent the likelihood of pixel regions being tampered with. The probability map output was then normalised to values between 0 and 1 across the dataset for each algorithm. The confusion matrix values are calculated using the normalised probability map of the corresponding images as the threshold. As shown in Table Table 1, where TR is the Tampered Result probability map value and AR is the Authentic Result probability map value, if the TR is more than or equal to the AR then it is deemed as tampered with. Then the actual mask is used to determine if this was a correct identification. Using the authentic image as the threshold in this experiment allows for the evaluation of differences across image classes and inpainting tools. However, it is important to note that this approach would not be possible in real life forensics scenarios; it is used here for the purpose of identifying key components and artefacts of images for the localisation of GenAI manipulation, particularly inpainting.

Table 1: Confusion Matrix value breakdown, where TR is the Tampered Result probability map value and AR is the Authentic Result probability map value

	Mask	Predicted
True Positive (TP)	True	TR >= AR
False Positive (FP)	False	TR >= AR
True Negative (TN)	False	TR < AR
False Negative (FN)	True	TR < AR

To evaluate the influence that the image classes and tools have on localisation methods, the data in the .mat files were organised with a MATLAB script that restructured them based on their class and tool fields. This allowed for the evaluation metrics to be calculated on the full dataset, as well as per class and per tool.

For the visual analysis, heatmaps were created using the MATLAB imagesc function, which creates a colour map from the values. This heatmap was saved to a pdf for all the images within each algorithm. The probability maps passed to this function were the binary threshold maps from the predicted function within Table 1. Full implementation details are available on GitHub[1].

3.3 Evaluation Metrics

The evaluation metrics selected for this experiment include the confusion matrix alongside the accuracy, precision, recall and F1-scores. The discussion will mainly focus on accuracy and F1-score due to their relevance in digital forensics investigations. Accuracy

[1]https://github.com/MatthewT0/GenAI-Image-Forensics-Toolbox

assesses the localisation methods' ability to correctly identify tampered and untampered areas, which is vital for ensuring reliability within digital forensics investigations. Using accuracy alone can provide misleading results, which is why the F1-score will also be used. This addresses accuracy limitations by balancing precision and recall, offering a robust metric for evaluating the performance in scenarios where both false positives and false negatives carry significant weight. This balance is crucial in digital forensics investigations, where undetected tampered images or images falsely identified to be tampered with when they are authentic could compromise the integrity of a case. By combining these metrics, we ensure a comprehensive evaluation of localisation methods.

4 Results and Discussion

The results from both experiments are displayed in three ways. A visual heatmap shows the highlighted tampered and untampered regions identified by the probability map. Secondly, a bar chart is used to represent similarities grouped by different dataset components, and finally a table presents all evaluation metrics for comparison purposes. For demonstration purposes, the heatmaps from the NOI4 algorithm were used due to having particularly strong results. These heatmaps are the binary decision values once the normalised probability maps from the tampered image and authentic reference image are compared. During the review process, a manual inspection of the output heatmaps was performed to identify images that could be detected through the human eye. These were selected based on obvious mask detection without looking at the authentic or tampered images.

4.1 Experiment 1: Image Classes

Experiment 1 explored the influence of the subject class on the performance of various localisation algorithms. The dataset was divided into four classes of animals, objects, persons, and scenery. The probability map from the animals subject class image once evaluated against the threshold probability map values can be seen in Figure 2. The figure is from the same example shown in Figure 1, where the cow is removed from the image. The heatmap highlights the detected tampered areas in yellow, where the cow is correctly identified and the sky is incorrectly identified as tampered. Through the visual heatmaps inspection there were notably more identifications in the animals image class, with 32% of all visual identifications being from this class.

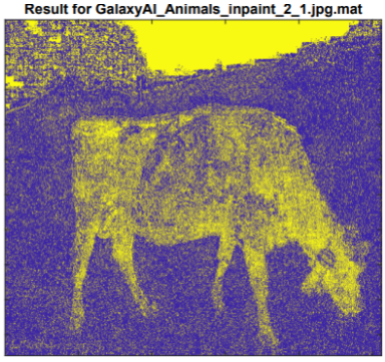

Figure 2: NOI4 localisation result, where yellow is tampered.

The F1-scores across classes are illustrated as a bar chart in Figure 3, where a consistent pattern can be identified between the animals and persons classes, achieving higher scores than objects or scenery across most algorithms. This suggests that the localisation methods are more effective in identifying the tampered regions within these two classes. In contrast, the objects and scenery classes show notably lower scores averaging at around half the other two classes' F1-scores, at 10% compared to 20%. The consistent underperformance in objects and scenery categories highlights a potential weakness in the handling of these image classes, indicating a need for future work. The F1 scores across the evaluated categories highlight the importance in having a diverse dataset of varying classes. However, to better determine which image attributes in the class cause such opposing F1-scores, a more in-depth analysis will need to be conducted.

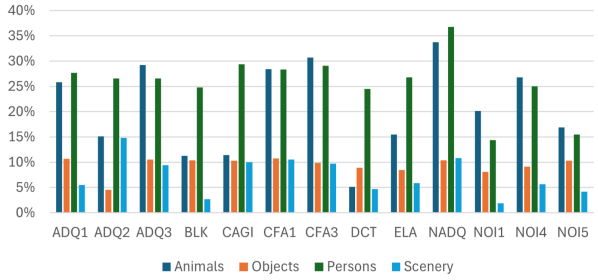

Figure 3: Average F1-score across algorithms, grouped by the class.

Table 2 presents the metrics for the full dataset as the baseline and the difference to each of the classes. The results highlight key trends in how the localisation methods perform across various image classes.

The objects class presents the highest TN and lowest FN results, demonstrating that the algorithms against this class are particularly effective in correctly identifying untampered areas within an image. In contrast, the precision and F1-score ratios decrease substantially, highlighting the difficulty of the classes in balancing the correct identifications of tampered regions with incorrect untampered regions. The results suggest that images within the objects class are more likely to be classified as untampered, regardless of whether they have been or not. Further investigation into the specific image attributes which cause the inaccuracy in image tampering detection must be conducted before a definitive analysis can be performed. However, it is important to note that other factors, such as smaller mask region, could be a result of inaccurate localisation.

On the contrary, the persons class can be seen to have the highest precision and F1-scores, but the lowest accuracy. This demonstrates the persons' class strength in correctly identifying tampered regions whilst struggling to identify the untampered areas correctly. The animals and persons classes share similar patterns in their F1-scores as previously highlighted, which is further reflected in Table 2. Both classes show confusion matrix values that deviate from the baseline in the same direction, increasing or decreasing together. This suggests that the two classes share similar underlying attributes that influence the localisation algorithms in comparable ways.

Table 2: Baseline metrics for the full dataset of tampered images compared to the per-class (60 images per class) differences from the baseline. Negative values indicate a decrease from the baseline metrics, while positive values represent an increase from the baseline metrics.

		Classes			
	Baseline	Animals	Objects	Persons	Scenery
TP	4.65	1.86	-2.67	3.75	-2.94
TN	55.13	-1.30	6.98	-8.98	3.30
FP	31.15	-5.03	1.28	-0.74	4.48
FN	9.06	4.47	-5.59	5.97	-4.84
Accuracy	59.79	0.56	4.31	-5.23	0.36
Precision	13.20	6.72	-7.16	8.36	-8.84
Recall	33.92	-1.44	2.48	1.92	-5.02
F1-Score	17.63	3.14	-8.22	8.16	-10.26

In summary, the results from the classes analysis highlight important patterns in how localisation methods perform across different image classes. This underscores the importance of a standardised and realistic testing environment through a set of clearly defined criteria. The similarities and patterns discovered between the persons and animals classes, compared to the objects and scenery classes, demonstrate the importance of rigorous testing to identify the features that influence the localisation performance. Establishing a set of criteria is essential for creating datasets that accurately reflect real-world scenarios, enabling the evaluation and creation of digital forensics techniques to overcome the rise of GenAI image manipulation.

4.2 Experiment 2: Tools

This experiment examines the impact of differing inpainting tools on the performance of localisation methods. The tools GLIDE, GalaxyAI, and Adobe Photoshop were selected for this evaluation due to their distinct approaches to tampering, which could introduce varying characteristics in tampered regions. The goal of this experiment is to determine whether the tool used affects the localisation ability, highlighting the importance of tool diversity in the creation of realistic datasets.

Across the tools utilised, a visual heatmap inspection was performed where Adobe Photoshop generally showed more obvious tampering indications, consisting of 44.62% out of identifications. However, some occasions were identifiable due to the mask edges being flagged as not tampered with and a large distribution of tampered indications falling within the mask region. Visual examples of manually identified tampering can be seen in Figure 4, where examples are shown of GLIDE, GalaxyAI and Photoshop respectively. Out of the manually identified masks, the GalaxyAI masks were more obvious to notice when they occurred due to more correct TN pixels being identified, but there were more Adobe Photoshop instances.

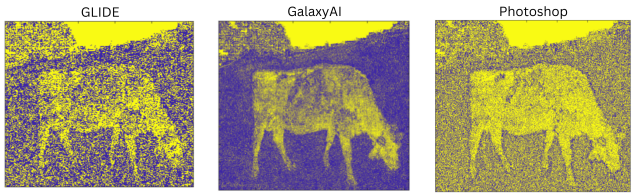

Figure 4: NOI4 visual heatmaps for the inpainting tools.

The accuracy results, as shown in Figure 5, reveal that GalaxyAI consistently demonstrates a higher performance compared to Photoshop and GLIDE's accuracy, averaging at 63%. This suggests that GalaxyAI tampered images have more detectable artefacts within the images, possibly due to their built-in watermarking. Photoshop and GLIDE's accuracy are generally lower across algorithms, demonstrating that these images are harder to detect. For GLIDE, the performance difference appears to be influenced by the lower resolution of GLIDE tampered images compared to the other tools, as the GLIDE code downscales the image resolution. The resolution disparity strongly suggests a contributing factor to the decreased localised performance. The underlying reason behind GalaxyAI's higher accuracy and GLIDE's lower accuracy would require further investigation to fully understand the contributing factors. In the case of Photoshop, the factors being its lower accuracy remain unclear and further experimentation would be beneficial in determining what components make these tampered images harder to detect.

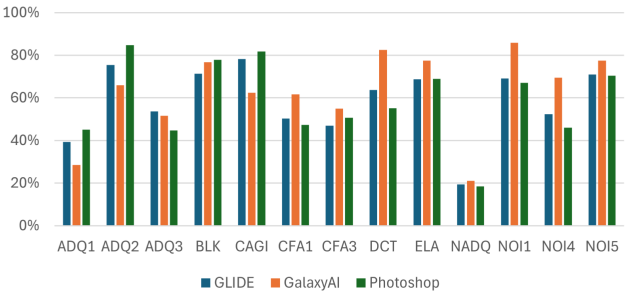

Figure 5: Average accuracy across algorithms, grouped by the tool.

The baseline metrics for the full dataset and the difference between ratios from the baseline to the classes can be seen in Table 3. GalaxyAI differentiates the most from the baseline, with the highest TN, accuracy, and precision, as well as the lowest FP. These results indicate that the localisation algorithms applied to GalaxyAI tampered images were generally more effective at correctly identifying untampered areas, averaging at 63% accuracy.
For all evaluation metrics except precision, GLIDE and Photoshop deviate from the baseline in the same direction. This consistent deviation in most metrics indicates similarities in the localisation difficulty of these tools.

Table 3: Baseline metrics for the full dataset of tampered images compared to the per-tool differences from the baseline. Negative values indicate a decrease from the baseline metrics, while positive values represent an increase from the baseline metrics.

		Tools		
	Baseline	GLIDE	GalaxyAI	Photoshop
TP	4.65	0.17	-0.26	0.10
TN	55.13	-1.57	3.19	-1.62
FP	31.15	1.56	-3.18	1.63
FN	9.06	-0.15	0.26	-0.11
Accuracy	59.79	-1.41	2.93	-1.52
Precision	13.20	-0.43	1.79	0.44
Recall	33.92	1.16	-1.91	0.75
F1-Score	17.63	-0.23	-1.04	-0.45

The conducted experiment builds upon the findings from the subject class analysis by focusing on the tools used to generate the tampered images rather than the content itself. From this, Photoshop provided the highest number of human eye detections from the output heatmaps, whilst GalaxyAI should have the most distinctive heatmaps. Additionally, GalaxyAI consistently demonstrated higher accuracy than the other two tools across algorithms, indicating more identifiable attributes in these images. Furthermore, GalaxyAI also showed the most significant deviations from the baseline with a 3.19% increase to TN and a 3.18% decrease from FP. This suggests that the localisation algorithms were typically producing more non-tampered classifications within their probability maps.

5 Conclusion

Many GenAI tampering datasets do not include a variety of manipulation types, classes or tools within them, creating a limited testing environment for digital forensics. This can result in the evaluation of detection and localisation techniques suffering from overfitting and generalisation, leading to their performance decreasing greatly when applied to real-world scenarios. An example of this is when GenAI localisation techniques are used in areas that differ from its original testing environment, such as a tampering technique that is used for detecting the alteration of people being used on identifying removed objects. This paper explored the need for standardised testing environments to enhance the performance of upcoming detection and localisation techniques being developed.

Addressing RQ1, it was found that certain classes had a higher level of influence on the performance of the localisation algorithms compared to others. Based upon the manual heatmap inspection, the animals class had more notable identifications, being 31.97% of the overall identifications. This class, along with the persons class, illustrated key patterns, consistently scoring higher F1-score by almost double the other classes analysed, at 20% compared to 10%. The identified similarities emphasise the importance of the inclusion of a variety of classes within the testing environment for digital forensics, where analysts often face a wide range of image classes from animals to objects.

The evaluation aimed at addressing RQ2 showed that Photoshop had the highest notable visual heatmaps, being 44% of the overall identifications. However, the GalaxyAI manual identifications were clearer with more correct TN values. Furthermore, GLIDE inpainted images were the most challenging for algorithms to detect, with consistently lower accuracy compared to GalaxyAI and Photoshop. This highlights the need for further attribute testing on the impact of lower resolution, which could be causing the lower accuracy for GLIDE inpainted images.

Having a realistic dataset that can cover a range of classes and tools will provide a robust testing environment for the development of techniques to counteract AI-based manipulation and cover the nature of media that forensic analysts handle. Without standardised criteria for datasets, the detection and localisation methods being developed cannot be accurately and rigorously tested to ensure their reliability in forensic investigations.

Additional components that should be further investigated for inpainting detection and localisation include the mask size, image complexity, and image resolution. These components alongside the underlying artefacts, such as watermarking in GalaxyAI, highlight the requirements for establishing a comprehensive set of criteria as raised in RQ3.

6 Future Work

The work conducted was performed at a small-scale level, and while it provided preliminary results and patterns, a larger dataset is necessary to validate these and establish the criteria for standardisation of datasets. Expanding the dataset would facilitate an extensive and comprehensive analysis of image attributes and dataset components.

Although a manual visual inspection of the heatmaps was conducted, future experimentation could include automating this process by analysing the pixel distribution of identified tampered areas, minimising any implication of human error. Furthermore, the analysis was conducted using a single threshold option, specifically the authentic image probability map. Whilst the data collected from the authentic image as the threshold provided interesting results, this is not feasible in real-world examples where the authentic image and prior knowledge are unknown. Therefore, additional threshold tests to determine if blind localisation is possible would be beneficial in improving forensic applicability. Further analysis into an error percentage margin around the threshold value should be investigated to determine if the threshold is suitable.

Lastly, additional testing on the localisation abilities of further file types, such as PNGS and WEBPs, would be invaluable in determining the various formats required for a diverse dataset. The MATLAB toolbox was specifically created for JPG images, although some algorithms allow for PNGs. However, MATLAB lacks official support for WEBP image format, so alternative solutions would be required to address this. File type testing may include various conversion approaches, including converting each of the three aforementioned file types into every other type to assess whether the direction of image conversion impacts the outcomes. Additionally, comparing double-compressed JPG images and WEBP images may provide insight to determine whether any additional components of the images are being identified from the localisation algorithms.

References

[1] Adobe. 2024. Adobe Firefly - Free Generative AI for creatives. https://www.adobe.com/uk/products/firefly.html

[2] Irene Amerini, Rudy Becarelli, Roberto Caldelli, and Andrea Del Mastio. 2014. Splicing forgeries localization through the use of first digit features. In *2014 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 143–148. https://doi.org/10.1109/WIFS.2014.7084318 ISSN: 2157-4774.

[3] Tiziano Bianchi, Alessia De Rosa, and Alessandro Piva. 2011. Improved DCT coefficient analysis for forgery localization in JPEG images. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2444–2447. https://doi.org/10.1109/ICASSP.2011.5946978 ISSN: 2379-190X.

[4] Tiziano Bianchi and Alessandro Piva. 2012. Image Forgery Localization via Block-Grained Analysis of JPEG Artifacts. *IEEE Transactions on Information Forensics and Security* 7, 3 (June 2012), 1003–1017. https://doi.org/10.1109/TIFS.2012.2187516 Conference Name: IEEE Transactions on Information Forensics and Security.

[5] Jordan J. Bird and Ahmad Lotfi. 2024. CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images. *IEEE Access* 12 (2024), 15642–15650. https://doi.org/10.1109/ACCESS.2024.3356122 Conference Name: IEEE Access.

[6] Ahmet Emir Dirik and Nasir Memon. 2009. Image tamper detection based on demosaicing artifacts. In *2009 16th IEEE International Conference on Image Processing (ICIP)*. IEEE, 1497–1500. https://doi.org/10.1109/ICIP.2009.5414611 ISSN: 2381-8549.

[7] Pasquale Ferrara, Tiziano Bianchi, Alessia De Rosa, and Alessandro Piva. 2012. Image Forgery Localization via Fine-Grained Analysis of CFA Artifacts. *IEEE Transactions on Information Forensics and Security* 7, 5 (Oct. 2012), 1566–1577. https://doi.org/10.1109/TIFS.2012.2202227 Conference Name: IEEE Transactions on Information Forensics and Security.

[8] Xiao Guo, Xiaohong Liu, Zhiyuan Ren, Steven Grosz, Iacopo Masi, and Xiaoming Liu. 2023. Hierarchical Fine-Grained Image Forgery Detection and Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3155–3165. https://openaccess.thecvf.com/content/CVPR2023/html/Guo_Hierarchical_Fine-Grained_Image_Forgery_Detection_and_Localization_CVPR_2023_paper.html

[9] Chryssanthi Iakovidou, Markos Zampoglou, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2018. Content-aware detection of JPEG grid inconsistencies for intuitive image forensics. *Journal of Visual Communication and Image Representation* 54 (July 2018), 155–170. https://doi.org/10.1016/j.jvcir.2018.05.011

[10] Neal Krawetz. 2007. A Picture's Worth: Digital Image Analysis and Forensics. In *A Picture's Worth: Digital Image Analysis and Forensics*. Black Hat USA, USA, 1 to 31. https://blackhat.com/presentations/bh-dc-08/Krawetz/Whitepaper/bh-dc-08-krawetz-WP.pdf

[11] Weihai Li, Yuan Yuan, and Nenghai Yu. 2009. Passive detection of doctored JPEG image via block artifact grid extraction. *Signal Processing* 89, 9 (Sept. 2009), 1821–1829. https://doi.org/10.1016/j.sigpro.2009.03.025

[12] Yuanman Li, Liangpei Hu, Li Dong, Haiwei Wu, Jinyu Tian, Jiantao Zhou, and Xia Li. 2024. Transformer-Based Image Inpainting Detection via Label Decoupling and Constrained Adversarial Training. *IEEE Transactions on Circuits and Systems for Video Technology* 34, 3 (March 2024), 1857–1872. https://doi.org/10.1109/TCSVT.2023.3299278 Conference Name: IEEE Transactions on Circuits and Systems for Video Technology.

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

[14] Zhouchen Lin, Junfeng He, Xiaoou Tang, and Chi-Keung Tang. 2009. Fast, automatic and fine-grained tampered JPEG image detection via DCT coefficient analysis. *Pattern Recognition* 42, 11 (Nov. 2009), 2492–2501. https://doi.org/10.1016/j.patcog.2009.03.019

[15] Babak Mahdian and Stanislav Saic. 2009. Using noise inconsistencies for blind image forensics. *Image and Vision Computing* 27, 10 (Sept. 2009), 1497–1503. https://doi.org/10.1016/j.imavis.2009.02.001

[16] Marie-Helen Maras and Alex Alexandrou. 2019. Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos. *The International Journal of Evidence & Proof* 23, 3 (July 2019), 255–262. https://doi.org/10.1177/1365712718807226 Publisher: SAGE Publications Ltd.

[17] Mekhail Mustak, Joni Salminen, Matti Mäntymäki, Arafat Rahman, and Yogesh K. Dwivedi. 2023. Deepfakes: Deceptions, mitigations, and opportunities. *Journal of Business Research* 154 (Jan. 2023), 113368. https://doi.org/10.1016/j.jbusres.2022.113368

[18] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. https://doi.org/10.48550/arXiv.2112.10741 arXiv:2112.10741.

[19] Yogesh Patel, Sudeep Tanwar, Pronaya Bhattacharya, Rajesh Gupta, Turki Alsuwian, Innocent Ewean Davidson, and Thokozile F. Mazibuko. 2023. An Improved Dense CNN Architecture for Deepfake Image Detection. *IEEE Access* 11 (2023), 22081–22095. https://doi.org/10.1109/ACCESS.2023.3251417 Conference Name: IEEE Access.

[20] Md Awsafur Rahman, Bishmoy Paul, Najibul Haque Sarker, Zaber Ibn Abdul Hakim, and Shaikh Anowarul Fattah. 2023. Artifact: A Large-Scale Dataset With Artificial And Factual Images For Generalizable And Robust Synthetic Image Detection. In *2023 IEEE International Conference on Image Processing (ICIP)*. 2200–2204. https://doi.org/10.1109/ICIP49359.2023.10222083

[21] Samsung. 2024. How to use Generative photo editing with Galaxy AI. https://www.samsung.com/sg/support/mobile-devices/how-to-use-generative-photo-editing-on-the-galaxy-s24/

[22] Maria-Paz Sandoval, Maria de Almeida Vau, John Solaas, and Luano Rodrigues. 2024. Threat of deepfakes to the criminal justice system: a systematic review. *Crime Science* 13, 1 (Nov. 2024), 41. https://doi.org/10.1186/s40163-024-00239-1

[23] Deependra Kumar Shukla, Abhishek Bansal, and Pawan Singh. 2024. A survey on digital image forensic methods based on blind forgery detection. *Multimedia Tools and Applications* (Jan. 2024). https://doi.org/10.1007/s11042-023-18090-y

[24] Satyendra Singh and Rajesh Kumar. 2024. Image forgery detection: comprehensive review of digital forensics approaches. *Journal of Computational Social Science* (April 2024). https://doi.org/10.1007/s42001-024-00265-8

[25] Shobhit Tyagi and Divakar Yadav. 2023. A detailed analysis of image and video forgery detection techniques. *The Visual Computer* 39, 3 (March 2023), 813–833. https://doi.org/10.1007/s00371-021-02347-4

[26] Jonas Wagner. 2015. Noise Analysis for Image Forensics. http://29a.ch/2015/08/21/noise-analysis-for-image-forensics

[27] Shuiming Ye, Qibin Sun, and Ee-Chien Chang. 2007. Detecting Digital Image Forgeries by Measuring Inconsistencies of Blocking Artifact. In *2007 IEEE International Conference on Multimedia and Expo*. IEEE, 12–15. https://doi.org/10.1109/ICME.2007.4284574 ISSN: 1945-788X.

[28] Markos Zampoglou, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2017. Large-scale evaluation of splicing localization algorithms for web images. *Multimedia Tools and Applications* 76, 4 (Feb. 2017), 4801–4834. https://doi.org/10.1007/s11042-016-3795-2

[29] Hui Zeng, Yifeng Zhan, Xiangui Kang, and Xiaodan Lin. 2017. Image splicing localization using PCA-based noise level estimation. *Multimedia Tools and Applications* 76, 4 (Feb. 2017), 4783–4799. https://doi.org/10.1007/s11042-016-3712-8

Received 25th November 2024; revised 13th January 2025; accepted 14th December 2024