Comprehensive literary review of detecting AI-generated and manipulated imagery

Comprehensive Literary Review of Detecting AI-generated and Manipulated Imagery An Analysis of Six Image Forensic Tools

Yu Leng
University of Guelph
Cybersecurity and Threat Intelligence
Guelph, Canada
leng@uoguelph.ca

Samuel Tracz
University of Guelph
Cybersecurity and Threat Intelligence
Guelph, Canada
traczs@uoguelph.ca

Xiaohai Wang
University of Guelph
Cybersecurity and Threat Intelligence
Guelph, Canada
xiaohai@uoguelph.ca

***Abstract*—Deepfake, as a groundbreaking technology, has leveraged powerful computational resources and cutting-edge algorithms to elevate visual experiences to unprecedented levels. However, behind this innovation lies a darker side: a growing number of malicious actors are exploiting this technology for harmful purposes. These include identity theft, portrait infringement, and the creation of fake images or videos used for extortion and misinformation. As a result, digital forensics is under increasing pressure to develop reliable and generalizable Deepfake detection mechanisms tailored to these challenges. In this study, we conduct a comparative analysis of six state-of-the-art open-source Deepfake image detection tools, along with the research papers they are based on. These methods are evaluated across various types of generative models, including GANs and diffusion models, with a focus on detection accuracy, robustness against image degradation, and generalization to unseen data. Our analysis highlights the practical strengths and limitations of each approach, offering insights for the deployment of future detection systems in real-world forensic workflows.**

***Index Terms*—Deepfake detection, AI-synthesized images, generative adversarial networks (GANs), diffusion models, digital forensics, open-source tools, media forensics, robustness evaluation.**

I. Introduction

Digital media has been transformed with the introduction of artificial intelligence enabling realistic and manipulated content. Deepfake, a mix of ”deep learning” and ”fake,” refers to AI-generated audio, video, and images that depict real individuals doing things they didn’t actually do [7] [8]. Although there may be benefits to this type of technology, deepfakes also show the potential to harm an individual. The field is quickly advancing with the use of models such as GANs and diffusion models, which can generate deepfakes without relying on pre-existing sources [9]. This introduces a need to detect such images to mitigate potential harm. Unlike traditional digital forensics, which focuses on recovering deleted data or analyzing hardware-based evidence, detecting deepfakes falls under the emerging branch of media forensics—concerned with identifying synthetic content and verifying digital authenticity. The detection of fake images has been practiced before the creation of deepfakes. However, just as how deep learning is improving the creation of fake images, it is also helping to detect it. The main methods of detection are General Visual Artifact Detection, GAN detection, and Diffusion detection. Deepfakes can introduce subtle clues that they are not genuine photos. Artifacts that are not in real images can be present and there can be irregularities in facial features that models introduce [10]. To counter these risks, the research community has proposed a variety of image forgery detection techniques aimed at identifying subtle differences between real and synthetic images. This study presents a comparative evaluation of six open-source deepfake image detection methods, each based on distinct detection strategies and architectural designs. These include: GAN Baseline [1] [11], HiFi-IFDL [3] [12], NPR [6] [13], GLFF [2] [14], DMDetector [4] [15] and CLIP-VIT [5] [16]. This paper systematically evaluates these methods in terms of detection accuracy, robustness to common distortions, and generalization to unseen forgery models. Through quantitative experiments and comparative analysis, we aim to provide practical insights into the strengths and limitations of each method for real-world deployment.

II. Detection Methods Overview

To support a comprehensive evaluation in this project, we selected six representative and publicly available deepfake image detection tools. These methods were chosen based on their diversity in detection strategies, model architectures, and relevance in recent literature. Prior to conducting the comparative analysis, we provide a detailed overview of each method, including its core technical characteristics, intended use cases, and operational procedures. This foundational understanding is essential for contextualizing the evaluations that follow. The comparisons will focus on several key aspects: detection accuracy across various generative models, robustness against post-processing operations such as compression and resizing, generalization to unseen data or generation techniques, and the scalability and practical usability of each method in real-world forensic workflows. These evaluation dimensions reflect the real-world challenges associated with deploying deepfake detection systems and enable a thorough assessment of each tool's effectiveness.

A. Gragnaniello et al. (2021): Baseline Evaluation

Gragnaniello et al. [1] systematically tested existing state-of-the-art GAN image detection methods to assess their performance and robustness under realistic conditions. Their study aimed to examine how well existing detectors generalize when faced with image distortions such as JPEG compression, downsampling, and previously unseen generative models. Seven detection approaches were analyzed, including models that analyze pixel patterns (Xception, Co-Net), frequency artifacts (Spec), and strategies to improve generalization (Wang2020, M-Gb). Training was conducted on 362,000 real and 362,000 synthetic images generated by ProGAN models, using 256×256 resolution. The evaluation was performed on a diverse set of GAN architectures not used in training, including StyleGAN, StyleGAN2, BigGAN, CycleGAN, StarGAN, ReIGAN, and GauGAN. High-resolution (1024×1024) datasets were also tested for broader assessment. Performance was measured using metrics such as detection accuracy and the ability to flag fake images while minimizing false alarms. Results showed that while most detectors performed well under ideal conditions (90%+ accuracy), their effectiveness dropped sharply when images were compressed or resized. The Wang2020 model which was used as a baseline outperformed other models in terms of robustness. A no-down variant of Wang2020 that avoids downsampling in early layers improved accuracy by 15% and detection rates by 14%, as it preserved fine details important for spotting synthetic artifacts. The study primarily emphasizes that the choice of architecture and the diversity of training data play a crucial role in building generalizable and robust deepfake detectors. It also highlights that detectors may struggle with inaccuracies caused by image distortions.

B. GLFF (Ju et al. 2022)

Ju et al. [2] propose the Global and Local Feature Fusion (GLFF), a deepfake image detector that integrates both global and local features to enhance generalization across diverse generative models. The method combines multi-scale global features which capture high-level semantics, and local artifacts which identify subtle inconsistencies through a two-branch architecture. The global branch uses ResNet-50 to extract hierarchical features, while the local branch employs a Patch Selection Module (PSM) to automatically identify informative regions using sliding windows (2×2 and 3×3). Selected patches are resized to 224×224 processed through the same ResNet-50, and fused with global features using a multi-head attention mechanism for final classification. The training dataset includes 362K real images from LSUN and 362K ProGAN-generated images, all at 256×256 resolution. The evaluation set comprises 128,424 images synthesized from 19 generation models, including GANS (StyleGAN3, BigGAN), transformer-based models, and edited datasets like Celeb-DF. These are divided into six model families for detailed benchmarking [14]. Experimental results show that GLFF outperforms several baselines on both seen and unseen generative models. In particular, it achieves higher mAP (mean Average Precision) on model families such as conditional GANs and diffusion models. The fusion of global and local cues significantly improves generalization, and ablation studies confirm the contribution of each component. This research highlights the importance of multi-level feature representation in deepfake detection and establishes a strong benchmark across a diverse set of synthetic image types.

C. HiFi-IFDL (Guo et al. 2023)

Guo et al. [3] propose HiFi-IFDL, a unified framework for both detecting and localizing forged images regardless of whether the forgery is fully synthesized by generative models or partially manipulated through traditional editing methods. The key innovation is the introduction of a hierarchical labeling system that classifies forgery types at multiple levels of granularity. These levels include distinctions between fully synthesized vs. partially manipulated content, the type of generation method (GAN, diffusion, or editing), conditional vs. unconditional models, and specific forgery techniques such as DDPM, STGAN, and Splicing. HiFi-IFDL uses a multi-branch architecture called HiFi-Net. Each branch is trained to classify forgery attributes at a specific level in the hierarchy. A color-frequency dual-path feature extractor is used to capture both spatial and frequency-domain artifacts. The network also has a pixel-wise localization module to generate binary masks indicating the manipulated regions in the image. The authors also construct a large-scale benchmark dataset named HiFi-IFDL, containing over 1.7 million training images, including both real images from datasets like FFHQ, AFHQ, CelebA-HQ, and MSCOCO, and forged images generated by 13 methods spanning GANs, diffusion models, and editing tools. Evaluation is performed on 7 public datasets, such as CASIA, Columbia, Coverage, and NIST16, as well as on diverse forgery sources including unseen domains. Quantitative results show that HiFi-IFDL achieves state-of-the-art performance in both detection (AUC up to 99.45) and localization tasks ($loU = 0.411$), outperforming other recent methods like Att.Xception and ObjectFormer. Ablation studies confirm the contribution of each architectural component, such as multi-branch learning and partial convolution. It demonstrates that combining hierarchical attribute learning with localization leads to improved robustness and fine-grained interpretability, making HiFi-IFDL a strong candidate for deployment in practical digital forensic scenarios.

D. NPR (Tan et al. 2023)

Tan et al. [6] introduce Neighboring Pixel Relationships (NPR), a deepfake detection method that leverages local pixel-level inconsistencies introduced by upsampling operations in generative models. Unlike conventional detectors that check global patterns or frequency-domain traces, NPR focuses on neighboring pixel relationships to reveal subtle artifacts that arise during image generation, especially in GANs and diffusion models. The model employs a lightweight CNN with only 1.44 million parameters, making it both efficient and deployable. The key idea is to extract local patches from input images and analyze their fine-grained texture consistency. The detector learns to recognize unnatural interpolation patterns caused by generative upsampling layers. During training, the authors use the ForenSynths dataset, comprising synthetic images generated by ProGAN and real images from LSUN. They also apply strong augmentations (e.g., Gaussian blur and JPEG compression) to improve robustness. Evaluation across 28 generative models shows the model's generalization capabilities. It achieves performance on traditional GANs (e.g., ProGAN, StyleGAN2, CycleGAN), diffusion models (e.g., DDPM, LDM, ADM, Stable Diffusion v1/v2), and autoregressive models (e.g., DALL-E). Additionally, real data from ImageNet, CelebA, FFHQ, and LAION datasets are used. The NPR detector achieves high accuracy on unseen models, showing strong generalization even when trained only on ProGAN. For example, it outperforms state-of-the-art methods like LGrad and Ojha by +7.1% accuracy on diffusion-generated images and +20.9% on challenging diffusion models like ADM and LDM. The authors emphasize that NPR works particularly well under limited supervision, making it suitable for real-world forensic workflows where training data is incomplete or biased. Through this experiment demonstrates that focusing on low-level structural patterns can significantly improve deepfake detection generalization rather than targeting global semantics, especially in cross-model scenarios.

E. DMimageDetection (Corvi et al. 2022)

Corvi et al. [4] address a pressing challenge in digital forensics: the detection of synthetic images generated by diffusion models (DMs), which are increasingly replacing GANs as the dominant paradigm for high-quality image generation. Unlike traditional detectors trained on GAN-generated content, this work demonstrates that many state-of-the-art classifiers fail to generalize to DM-generated images due to their unique low-level statistical properties. The authors test existing detectors—including CNNs and frequency-based models—on a comprehensive benchmark of images generated by modern diffusion models such as DDPM, LDM, and GLIDE. They observe that traditional GAN-trained detectors show significantly lower accuracy and precision when applied to diffusion-generated content, especially under realistic transformations such as JPEG compression and resizing. To address this, the authors propose training a model directly on DM-generated images using similar architectural setups to prior GAN-based detectors (e.g., ResNet-50). The evaluation uses a diverse set of real images (ImageNet, LSUN, CelebA-HQ) and synthetic images generated by ProGAN and various DMs, with an emphasis on unseen test cases. Experimental results show that while incorporating DM data in training improves performance on similar models, cross-model generalization still remains weak. Overall, this study highlights the limitations of existing detection tools when confronted with diffusion models, and provides one of the earliest systematic benchmarks for this emerging threat. This study highlights the need for new forensic detectors designed specifically for DM and proposes basic strategies for building such detectors.

F. CLIP-VIT (Ojha et al. 2024)

A novel approach to deepfake image detection is introduced by Ojha et al. [5] by leveraging the vision-language capabilities of the CLIP (Contrastive Language-Image Pretraining) model, particularly the CLIP-VIT backbone. Unlike traditional methods that require retraining classifiers on synthetic image datasets, this method explores zero-shot or low-shot detection by using the CLIP feature space for real/fake classification. The method is based on extracting feature embeddings using a fixed CLIP-VIT model and applying lightweight classifiers—such as nearest-neighbor (NN) search or linear classifiers (LC)—to distinguish between real and fake images. The authors evaluate their method across 28 generative models, including GANs (e.g., ProGAN, StyleGAN), diffusion models (e.g., LDM, Glide, DDPM), and autoregressive models (e.g., DALL-E mini). The evaluation dataset includes real images from ImageNet, LSUN, and LAION, and synthetic images generated under different diffusion steps and model variants. Key results show that CLIP-VIT, even without fine-tuning, achieves strong generalization, outperforming fully trained CNN baselines on unseen generative models. Specifically, it achieves a +9.8 mAP gain overall, and up to +19.49 mAP on unseen diffusion and autoregressive models like LDM and DALL-E, compared to ResNet-based classifiers trained from scratch The authors argue that the semantic embedding space of CLIP inherently captures high-level visual inconsistencies in synthetic images, enabling robust detection without retraining. This approach offers a scalable, low-resource alternative to traditional supervised deepfake detectors, particularly valuable in rapidly evolving generative model landscapes.

III. Methodology

This section details the experimental framework and procedures used to compare the six open-source deepfake image detection tools. Our design ensures a fair comparison by evaluating all methods under unified conditions, covering dataset construction, evaluation metrics, experimental design, and tool setup.

A. Dataset Construction

The composition of datasets were a critical component of the methodologies reported in the papers reviewed. A common approach was having a comprehensive dataset consisting of Real Image Datasets. High-quality, large scale, and publicly available image datasets such as FFHQ, CelebA, and LSUN, which provide a broad range of real face and scene images. Additionally, synthetic image sources were used in the reviewed papers and tools including GAN-generated images (e.g., ProGAN, StyleGAN, StyleGAN2, and BigGAN), diffusion model-generated images (e.g., DDPM, LDM, and GLIDE), and hybrid and edited samples created using a combination of generative models and traditional editing techniques. Furthermore, to evaluate cross-model generalization, we construct a ”cross-model test set” by excluding certain generative models' data during training. This set is used exclusively during testing to assess the ability of each method to detect unseen forgery types.

B. *Evaluation Metrics*

The authors use a variety of quantitative metrics to assess the efficacy of the detection tools. Accuracy and F1-Score measure the overall classification performance, with the F1-Score balancing precision and recall when distinguishing between real and synthetic images. Additionally, Average Precision (AP) integrates precision-recall trade-offs for a nuanced understanding of detection accuracy. The area under the receiver operating characteristic curve (AUC-ROC) is also employed to evaluate the model's ability to discriminate between classes at different thresholds, which is particularly crucial when dealing with imbalanced datasets. Robustness is evaluated by measuring performance under adversarial conditions such as image compression, resizing, and the addition of noise.

C. *Experimental Design*

Our experiments are structured as follows:

- **Unified Test Set:** All detection tools are evaluated on the same test set, which includes original images, compressed versions, resized images, and images with added noise. This ensures that performance comparisons are meaningful across diverse real-world scenarios.
- **Cross-Model Testing:** By excluding certain generative model outputs from the training phase, we test the detectors on previously unseen forgery types, thereby assessing their generalization capabilities.
- **Grouped Experiments:** We conduct separate experiments for different image types (e.g., GAN-generated vs. diffusion-generated) and report performance metrics for each group.
- **Repetition and Averaging:** Each experiment is repeated multiple times to minimize random variation, and the average values are reported as final metrics.

D. *Tool Setup and Experimental Environment*

The GitHub repositories that accompany the research papers each have unique configurations, as detailed in their respective README files. Due to technical limitations, we have decided to use the benchmark data instead of running the tools ourselves, since there are benchmarks available for all these tools.

IV. Comparative Analysis

This section presents our integrated evaluation of six deepfake detection tools across multiple dimensions, including detection performance on various generators, robustness to degradations, generalization to unseen models, and practical deployment considerations. The results are supported by quantitative metrics (AUC scores, inference speed,etc.) and visualizations (ROC curves and radar charts).

A. *Performance Comparison*

We evaluated each tool's benchmarks which are listed in the Appendix and summarized it into metrics AUC, Accuracy, F1-score and mAP across all the tools. Missing metrics are marked as N/A as some of the metrics were not found for those tools. Table I shows the summary of metrics.

B. *Critical Evaluation by Dimension*

1) Detection Accuracy: Detection accuracy varies significantly across the model types. GAN-focused tools such as GAN Baseline [1] achieves high accuracy on GAN-generated content, but seems to fail on diffusion models. Similarly, GLFF [2] excels in face forgery detection but struggles with diffusion content, especially if it doesn't contain a face. Cross-model tools such as Clip-ViT [5] perform better than supervised CNNs on diverse generative models by using certain pre-training for better detection. HiFi-IFDL [3] has a balanced accuracy and localization, which makes it the most suitable for forensic workflows that include multiple tasks.

2) Robustness to Image Degradation: Robustness is usually tested with compression, resizing and added noise. Most tools showed that they are sensitive to image distortions. All the models showed that their AUC or accuracy dropped under these conditions. However, some showed slightly lower drops with HiFi and NPR showing the most resilience.

3) Generalization to Unseen Models: The significant difference in performance between tools indicates that cross-model generalization is a substantial obstacle. For GAN and diffusion models, NPR generalizes the best, whereas GLFF fails on latent diffusion models. CLIP-VIT requires no fine-tuning to achieve high results.

4) Practical Usability: Deciding which tool is the most feasible is dependent on the computational demands. NPR is the most lightweight tool with 1.44 million parameters that enable real-time detection. HiFi-IFDL and CLIP-VIT seem to require significant GPU resources since they have been tested on higher-end GPUs. The tools also provide specialized uses. DMDetector provides insights into diffusion artifacts, but is not very adaptable to cross-models.

C. *Trends and Limitations*

The key trend is that there is a tradeoff between specialization and generalization. Tools like GLFF and HiFi-IFDL are

TABLE I
Summary of Tool Benchmarks

Tool	AUC	Accuracy	F1-Score	MAP	Highlights
GAN Baseline	∼90%	50%	N/A	N/A	High performance on specific GAN architectures, but poor generalization to diffusion models.
GLFF	85.1%	24.7%	N/A	N/A	Robust to unseen generation models and post-processing effects on face forgeries. Specific metrics need to be obtained from the paper.
HiFi-IFDL	N/A	84.3%	86.8%	N/A	Strong performance in both detection and localization across various forgery types and robust to post-processing.
NPR	N/A	71.7%	72.6%	96.9%	Good generalization across GANS and diffusion models on academic benchmarks, but challenges with in-the-wild deepfakes.
DMDetector	77.9%	60.4%	N/A	N/A	Early work focused on understanding diffusion model artifacts.
CLIP-VIT	N/A	95.09%	N/A	N/A	Excellent generalization to new generative models without explicit training on real/fake images.

very good at performing niche tasks, whereas CLIP-VIT prioritizes broader detection. This may mean that specialized tools are good for specific and well-defined scenarios but may lack versatility when faced with new types of image manipulation. On the other hand, tools designed for generalization may sacrifice some precision to offer greater robustness and adaptability across evolving threats. Therefore, the choice of tool depends heavily on the intended use case. Diffusion-generated content is a relatively new development with deepfakes, and presents challenges for current deepfake detection models. Many of the existing tools and techniques were primarily designed to detect GAN-generated content and struggle to generalize effectively to the distinctive characteristics of diffusion-generated images. This performance gap highlights a crucial gap in the current deepfake detection tools that needs to be focused on to ensure robust protection against AI-generated images. The academic benchmarks set for some of these tools are significantly different compared to when the tools perform in the real world. The benchmark data collected in this review tries to gather from both, and may not have true benchmark values due to limited data points.

D. *Recommendations for Deployment*

Table II describes the tool's strengths and the scenarios for which they are most suited.

V. Discussion

The analysis of deepfake detection tools highlights the challenge of balancing specialization and adaptability. Tools like the GAN Baseline excel at identifying images from specific models, such as ProGAN or StyleGAN, but struggle when faced with newer methods like diffusion models. This shows how quickly detection methods can become outdated as AI evolves. A major hurdle for these tools is handling real-world conditions. Most detectors perform well in controlled experiments but degrade significantly when images are compressed, resized, or changed in any way which is a common scenario on the internet. As the standard for generating deepfakes is increasing, detection tools need a way to keep up. The most practical approach would be to combine multiple techniques such that you get to take advantage of the specialized results of each tool. No matter the approach, the stakes are high as people start to lose trust in digital media as the rise of threats such as scams, misinformation, and identity theft grow.

TABLE II
TOOL RECOMMENDATIONS

Application	**Tool**	**Reason**
GAN-specific Detection	GAN Baseline	Optimized specifically for ProGAN/StyleGAN.
Multiple forgery types with localization	HiFi-IFDL	Robust detection and specific localization.
Cross-model generalization	CLIP-VIT	zero-shot flexibility.
Diffusion model analysis	DMDetector	Combines artifact detection techniques.
Face Forgery detection	GLFF	Best AUC on face forgeries.

VI. CONCLUSION AND FUTURE WORK

This study evaluated six deepfake detection research papers along with their tools, examining their strengths, weaknesses, and practical applications. While tools such as CLIP-VIT and NPR demonstrated potential for generalization across different generative models, they showed limitations in specific tasks (e.g., face forgery localization) and robustness against common image distortions. On the other hand, specialized detectors like GLFF excelled in targeted areas but failed to adapt to emerging threats like diffusion models. These findings show the critical need for hybrid detection methodologies and enhanced benchmark datasets to effectively counter the evolving landscape of AI-generated media and mitigate associated societal harms. The rapid evolution of generative AI has underscored the critical need for continuous innovation in detection technologies. To mitigate the potential for widespread harm caused by deepfakes, the development of lightweight hybrid detection systems and the implementation of real-world benchmarking are essential strategies for deepfake detection tools to keep pace with this evolving threat landscape.

REFERENCES

[1] D. Gragnaniello, D. Cozzolino, F. Marra, G. Poggi, and L. Verdoliva, ”Are gan generated images easy to detect? a critical analysis of the state-of-the-art,” in ICME. IEEE, 2021, pp. 1-6.

[2] Y. Ju, S. Jia, J. Cai, H. Guan, and S. Lyu, ”Giff: Global and local feature fusion for ai-synthesized image detection,” IEEE Transactions on Multimedia, 2023.

[3] X. Guo, X. Liu et al., ”Hierarchical fine-grained image forgery detection and localization,” in CVPR, 2023, pp. 3155-3165.

[4] R. Corvi, D. Cozzolino et al., ”On the detection of synthetic images generated by diffusion models,” in ICASSP. IEEE, 2023, pp. 1-5.

[5] U. Ojha, Y. Li, and Y. J. Lee, ”Towards universal fake image detectors that generalize across generative models,” in CVPR, 2023, pp. 24 480-24 489.

[6] C. Tan, Y. Zhao, S. Wei, G. Gu, P. Liu, and Y. Wei, ”Rethinking the upsampling operations in cnn-based generative network for generalizable deepfake detection,” arXiv preprint arXiv:2312.10461, 2023.

[7] M. Westerlund, ”The Emergence of Deepfake Technology: A Review,” ResearchGate, Nov. 2019. https://www.researchgate.net/publication /337644519_The_Emergence_of_Deepfake_Technology_A_Review.

[8] S. Gorkhover, ”Spotting the Deepfake,” IEEE Transmitter, Jul. 25, 2024. https://transmitter.ieee.org/spotting-the-deepfake/.

[9] Reza Babaei, S. Cheng, R. Duan, and S. Zhao, ”Generative Artificial Intelligence and the Evolving Challenge of Deepfake Detection: A Systematic Analysis,” Journal of Sensor and Actuator Networks, vol. 14, no. 1, pp. 17-17, Feb. 2025, doi: https://doi.org/10.3390/jsan14010017.

[10] H. Lee, C. Lee, K. Farhat, L. Qiu, S. Geluso, A. Kim, O. Etzioni, ”The tug-of-war between Deepfake Generation and detection,” arxiv.org, https://arxiv.org/html/2407.06174v4.

[11] grip-unina, ”GitHub - grip-unina/GANimageDetection,” GitHub, 2021. https://github.com/grip-unina/GANimageDetection (accessed Mar. 29, 2025).

[12] CHELSEA234, ”GitHub - CHELSEA234/HiFi_IFDL: Hierarchical Fine-Grained Image Forgery Detection and Localization (CVPR2023 and IJCV2024),” GitHub, 2023. https://github.com/CHELSEA234/HiFi_IFDL.

[13] chuangchuangtan, ”GitHub - chuangchuangtan/NPR-Deepfake Detection,” GitHub, 2023. https://github.com/chuangchuangtan/NPR-Deepfake Detection.

[14] littlejuyan, ”GitHub - littlejuyan/Fusing GlobalandLocal,” GitHub, 2022. https://github.com/littlejuyan/Fusing GlobalandLocal.

[15] grip-unina, ”GitHub - grip-unina/DMimage Detection: On the detection of synthetic images generated by diffusion models,” GitHub, 2022. https://github.com/grip-unina/DMimage Detection.

[16] WisconsinAlVision, ”GitHub - WisconsinAlVision/Universal FakeDetect,” GitHub, 2023. https://github.com/Yuheng-Li/Universal Fake Detect.

[17] NVlabs, ”GitHub - NVlabs/stylegan3-detector,” GitHub, 2021. https://github.com/NVlabs/stylegan3-detector.

[18] M. Mulki, S. Mulki, ”Spotting Diffusion: Using transfer learning to detect Latent Diffusion Model-synthesized images — Journal of Emerging Investigators,” Emerginginvestigators.org, 2024. https://emerginginvestigators.org/articles/23-256.

[19] X. Guo, X. Liu, I. Masi, and X. Liu, ”Hierarchical Fine-Grained Image Forgery Detection and Localization: Supplementary material” International Journal of Computer Vision, Dec. 2024, doi: https://doi.org/10.1007/s11263-024-02255-9. https://cvlab.cse.msu.edu/pdfs/guo_liu_ren_grosz_masi_liu_cvpr2023_supp.pdf.

[20] BitMind, ”Deepfake Detection Arena Leaderboard,” Huggingface.co, 2025. https://huggingface.co/spaces/bitmind/dfd-arena-leaderboard.

[21] S. Jia, et al. ”Can ChatGPT Detect DeepFakes? A Study of Using Multimodal Large Language Models for Media Forensics,” arxiv.org. https://arxiv.org/html/2403.14077v1.

[22] J. Xu, Y. Yang, H. Fang, H. Liu, W. Zhang, ”FAMSEC: A Few-shot-sample-based General AI-generated Image Detection Method,” Arxiv.org, 2021. https://arxiv.org/html/2410.13156v1.

Appendix

TABLE III
BENCHMARK RESULTS FOR GAN BASELINE

Dataset	Forgery Type	Metric	Value	Notes
ProGAN images	GAN	AUC	0.95 [17]	
ProGAN/StyleGAN 2 images	GAN	AUC	0.95 [17]	
StyleGAN3 FFHQ-U (no compression)	GAN	AUC	0.97 [17]	20K images
StyleGAN3 FFHQ-U (no compression)	GAN	AUC	0.95 [17]	20K images
StyleGAN3 FFHQ-U (with compression)	GAN	AUC	∼0.82 [17]	-15% decrease from no compression
Various low-resolution GAN-generated images	GAN	AUC	>0.9 [1]	no-down
Various high-resolution GAN-generated images	GAN	AUC	>0.95 [1]	Behavior similar to low-resolution; baseline accuracy better
ADM-generated images (retrained)	Diffusion Model	Accuracy	0.51 [18]	
Testing subset (diffusion models)	Diffusion Model	Accuracy	0.50 [18]	ResNet50-NoDown-StyleGAN2

TABLE IV
BENCHMARK RESULTS FOR HIFI

Dataset	Task	Metric	Value	Notes
HiFi-IFDL	Detection	F1	86.8% [3]	Overall
HiFi-IFDL	Detection	F1	81.5% [3]	CNN synthesis
HiFi-IFDL	Detection	F1	88.2% [3]	Image Editing
NIST16	IFDL	Accuracy	84.3% [19]	Generally competitive or better than SPAN, PSCC, Obj.Fo.

TABLE V
NPR BENCHMARK RESULTS

Dataset	Forgery Type	Metric	Value	Notes
AIGCDetectBenchmark	Mixed Deepfake Methods	Average Accuracy	91.7% [13]	Using ProGAN-4class checkpoint
GenImage	Al-Synthesized	Mean Accuracy	90.1% [13]	Across various GANs and diffusion models
GenImage	Al-Synthesized	Mean AP	96.9% [13]	Across various GANs and diffusion models
DFD-Arena	Al-Generated	Accuracy	0.7169 [20]	
DFD-Arena	Al-Generated	Precision	0.9193 [20]	
DFD-Arena	Al-Generated	Recall	0.5996 [20]	
DFD-Arena	Al-Generated	F1-Score	0.7258 [20]	
DFD-Arena	Al-Generated	MCC	0.5044 [20]	
DFD-Arena (CelebA-HQ)	Deepfake	Accuracy	0.987 [20]	
DFD-Arena (Flickr30k)	Al-Generated	Accuracy	0.916 [20]	
DFD-Arena (ImageNet)	Al-Generated	Accuracy	0.834 [20]	
DFD-Arena (DiffusionDB)	Diffusion Model	Accuracy	0.876 [20]	
DFD-Arena (CelebA-HQ-SDXL D)	Diffusion Model	Accuracy	0.386 [20]	
DFD-Arena (CelebA-HQ-Flux)	Diffusion Model	Accuracy	0.846 [20]	
DFD-Arena (Flickr30k-SDXL)	Diffusion Model	Accuracy	0.302 [20]	
DFD-Arena (MS-COCO-Flux)	Diffusion Model	Accuracy	0.588 [20]	

TABLE VI
BENCHMARK RESULTS FOR GLFF

Dataset	Forgery Type	Metric	Value	Notes
DF^3 Dataset	Unprocessed, Common Post-processing, Face Blending, Anti-Forensics, Multi-image Compression, Mixed	AUC	0.813	Superior performance reported. Specific metrics need to be obtained from the paper [2].
Real (FFHQ) vs. Synthetic (StyleGAN2)	Face Forgery	Accuracy	82.9% [21]	Raw data. Nodown, BeyondtheSpectrum
Real (FFHQ) vs. Synthetic (Latent Diffusion)	AI-Synthesized	Accuracy	0.2% [21]	Raw data. This unusually low value requires further investigation in the original paper. Nodown, BeyondtheSpectrum
Real (FFHQ) vs. Synthetic (StyleGAN2)	Face Forgery	AUC	97.5% [21]	Raw data. Nodown, BeyondtheSpectrum
Real (FFHQ) vs. Synthetic (Latent Diffusion)	AI-Synthesized	AUC	86.7% [21]	Raw data. Nodown, BeyondtheSpectrum
Real (FFHQ) vs. Synthetic (StyleGAN2)	Face Forgery	Accuracy	7.6% [21]	Post-processed. This low value requires further investigation in the original paper. Nodown, BeyondtheSpectrum
Real (FFHQ) vs. Synthetic (Latent Diffusion)	AI-Synthesized	Accuracy	8.1% [21]	Post-processed. This low value requires further investigation in the original paper. Nodown, BeyondtheSpectrum
Real (FFHQ) vs. Synthetic (StyleGAN2)	Face Forgery	AUC	80.6% [21]	Post-processed. Nodown, BeyondtheSpectrum
Real (FFHQ) vs. Synthetic (Latent Diffusion)	AI-Synthesized	AUC	79.4% [21]	Post-processed. Nodown, BeyondtheSpectrum

TABLE VII
BENCHMARK RESULTS FOR DMDETECTOR [4]

Model	**Spec Uncomp. Acc./AUC%**	**PatchFor. Uncomp. Acc./AUC%**	**Wang2020 Uncomp. Acc./AUC%**	**Grag2021 Uncomp. Acc./AUC%**	**Spec Resized & Comp. Acc./AUC%**	**PatchFor. Resized & Comp. Acc./AUC%**	**Wang2020 Resized & Comp. Acc./AUC%**	**Grag2021 Resized & Comp. Acc./AUC%**
ProGAN	83.5/99.2	64.9/97.6	99.9/100	99.9/100	49.7/48.5	50.4/65.3	99.7/100	99.9/100
StyleGAN2	65.3/72.0	50.2/88.3	74.0/97.3	98.1/99.9	51.8/50.5	50.8/73.6	54.8/85.0	63.3/94.8
StyleGAN3	33.8/4.4	50.0/91.8	58.3/95.1	91.2/99.5	52.9/51.9	50.2/76.7	54.3/86.4	58.3/94.4
BigGAN	73.3/80.5	52.5/85.7	66.3/94.4	95.6/99.1	52.1/52.2	50.5/58.8	55.4/85.9	79.0/99.1
EG3D	80.3/89.6	50.0/78.4	59.2/96.7	99.4/100	58.9/60.6	49.8/81.9	52.1/85.1	56.8/96.6
Taming Tran.	79.6/86.6	50.5/69.4	51.2/66.5	73.5/96.6	49.0/49.1	50.0/64.1	50.5/71.0	56.2/94.3
DALL E Mini	80.1/88.1	51.5/82.2	51.7/60.6	70.4/95.6	59.1/61.9	50.1/68.7	51.1/66.2	62.3/95.4
DALL-E 2	82.1/93.3	50.0/52.5	50.3/85.8	51.9/94.9	62.0/65.0	49.7/58.4	50.0/44.8	50.0/64.4
GLIDE	73.4/81.9	50.3/96.6	51.1/62.6	58.6/86.4	53.1/52.5	51.0/71.5	50.3/65.9	51.8/90.0
Latent Diff.	72.1/78.5	51.8/84.3	51.0/62.5	58.2/91.5	47.9/46.3	50.6/65.2	50.7/69.1	52.4/89.4
Stable Diff.	66.8/74.7	50.8/85.0	50.9/65.9	62.1/92.9	46.5/44.5	51.1/77.2	50.7/72.9	58.1/93.7
ADM	55.1/53.3	50.4/87.1	50.6/56.3	51.2/57.4	49.1/49.1	51.0/69.1	50.3/68.1	50.6/77.2
AVG	70.5/75.2	51.9/83.2	59.5/78.6	75.8/92.8	52.7/52.7	50.4/69.2	55.8/75.0	61.5/90.8

TABLE VIII
BENCHMARK RESULTS FOR CLIP-VIT

Dataset	**Forgery Type**	**Metric**	**Value**	**Notes**
Unseen diffusion and autoregressive models	AI-Generated	mAP Improvement	+15.07 [5]	”State-of-the-Art generalization performance” [5]
Unseen diffusion and autoregressive models	AI-Generated	Accuracy Improvement	+25.90% [5]	
ForenSynths, Universal FakeDetect, GenImage	forgery awareness module (FAM) and semantic feature-guided contrastive learning (SeC)	Accuracy Average	95.09% [22]	Using FAMSEC (enhancement of CLIP-VIT)