Improving SPAI's Robustness with Challenging Datasets

David Werkhoven*, **Jasper van der Valk***, **Rick van der Veen***, **Xin Yu Zhu***
Informatics Institute
University of Amsterdam
Amsterdam, XH 1098
{david.werkhoven, jasper.van.der.valk, rick.van.der.veen, xinyu.zhu2}
@student.uva.nl

Abstract

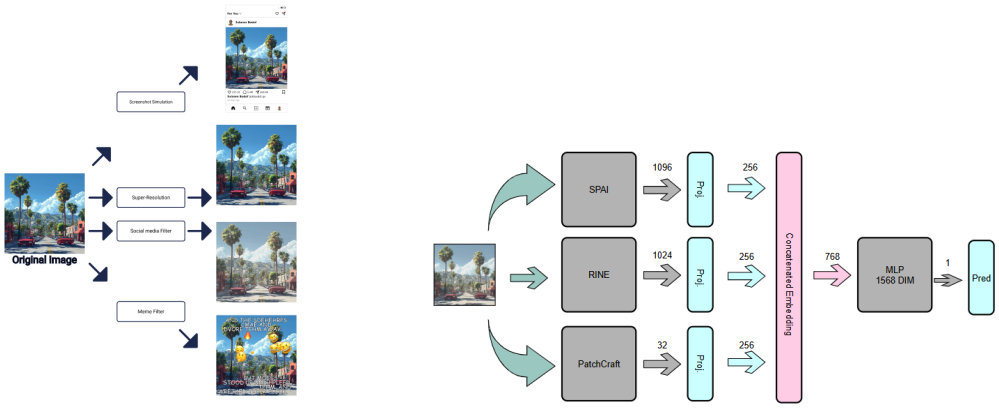
AI-generated images are often shared in real-world scenarios through memes, filters, or social media, where they are modified into derivative images. Existing AI-generated image detection (AID) methods, such as SPAI, struggle to detect these altered images. We introduce a dataset pipeline with four robustness tests that simulate real-world image modifications and evaluate SPAI on these challenging scenarios. Results show a significant drop in SPAI's performance. To address this, we propose ROGER, a multi-modal model combining SPAI with complementary techniques, achieving improved detection accuracy across all test cases. The code is publicly available at. https://github.com/Rickvanderveen/deep-learning-2.git

1 Introduction

Generative models in artificial intelligence (AI) have made significant progress in generating photo-realistic images [15, 9, 30]. This introduces new problems such as fake content being indiscernible from real content, which poses a real danger to both individuals and the public sphere [24] and, as such, calls for AI-generated image detection (AID). While prior research has developed detection methods, these are usually specialized for particular generators and tend to fail to generalize across different or unseen generators [15]. Although more recent methods have started to move away from this approach, they often still struggle to detect images in real-life scenarios [15], such as AI-generated images that are uploaded to social media, edited with a filter, or pasted into a meme. This is because current AID methods primarily focus on a narrow set of image features or specific patterns within the data, rather than capturing the broader, underlying structures of an image. This regularly leads to these detection methods failing to accurately identify the image as AI-generated. [15]. Building on this concern, this report extends the work of the SPectral AI-generated Image detection (SPAI) paper [15] by further investigating methods for detecting AI-generated images. SPAI was selected as it is the current state-of-the-art and best-performing model for AID. However, as it relies solely on the spectral distribution of an image, it is not robust when the spectral distribution is distorted in certain ways [15]. As such, this report makes three contributions. First, it introduces a pipeline to create four datasets that simulate real-world modifications of AI-generated images, referred to as derivative images (Figure 1a). Second, it utilizes these datasets of derivative images to evaluate the robustness of SPAI. Third, it combines the original SPAI approach with the techniques of two additional papers, RINE [16] and PatchCraft [33] (Figure 1b), to investigate the development of the RObust AI-GEnerated image Recognizer (ROGER), which is a more generalized and robust detection model. Using ROGER for AID, the performance with respect to SPAI increases significantly, both with modified and unmodified datasets.

*Equal contribution

37th Conference on Neural Information Processing Systems (NeurIPS 2023).



(a) Pipeline of making the modified datasets

(b) Pipeline of ROGER

Figure 1: Visual representation of both the derivative image pipeline and the integrated model approach pipeline.

2 Related work

Earlier detection methods have attempted to learn the common spatial artifacts in generated images [13] for specific image generators. However, since each generator produces different artifacts, it is impossible to learn all of them. This challenge further hinders AID methods, as they struggle to generalize to newly created generators [13, 20, 15]. To address this limitation, only the characteristics of real images are learned, enabling AID methods to generalize to unseen generators [20, 15, 23]. Earlier work targeted Generative Adversarial Network (GAN)-based generators, but newer diffusion models produce more diverse and realistic images [8, 5, 25, 21]. The artifact differences between GANs and diffusion models are even more evident than those among models of the same type [23]. Newer methods make use of the frequency domain of an image, which has been shown to provide valuable information for AID [20, 15]. The continuous development of image generation models can cause AID to perform poorly on newer models. Methods such as DMID [5] and RINE [16] are capable at AID but fail to generalize across the latest image generators [15]. The SPAI method generalizes the best across older and latest image generators [15]. It classifies AI-generated images as out-of-distribution by learning the spectral distribution of real images. It outperforms state-of-the-art methods, is more robust to perturbation, and works with images of any resolution [15]. However, SPAI relies solely on spectral information and struggles to identify AI-generated images when the spectral distribution is distorted.

3 Reproducibility of SPAI

Since this work builds upon SPAI, its performance is reproduced to verify its reproducibility and results. This is done using the author's implementation and pretrained weights of SPAI. Model predictions are evaluated on the same metric by computing the average AUC for each fake image test set over five real image test sets. The reproduced results are presented in Table 1. It shows that the reproduced results do not differ much from the original results, making SPAI [15] and its framework highly reproducible.

Table 1: Comparison of SPAI with the reproduced SPAI results. Reported is the AUC averaged over five datasets of real images. The fake datasets used are show in Table 1. The respective test sets for real images are ImageNet [7], COCO [19], OpenImages [17], FODB [12], and RAISE [6].

Image Size	<0.5 MPixels			0.5 - 1.0 MPixels						>1.0 MPixels				**AVG**
Approach	Glide [22]	SD1.3 [1]	SD1.4 [1]	Flux [18]	DALLE2 [27]	SD2 [1]	SDXL [26]	SD3 [10]	GigaGAN [14]	MJv5 [1]	MJv6.1 [28]	DALLE3 [2]	Firefly [1]	
SPAI [15]	90.2	99.6	99.6	83.0	91.1	96.5	97.4	75.9	85.4	94.5	84.0	90.2	96.0	**91.0**
SPAI reproduced	90.5	99.6	99.6	83.5	91.5	96.6	97.5	76.5	85.8	94.7	84.4	90.6	96.1	**91.3**

4 Extension

4.1 Image modifications

To assess the robustness of SPAI, four new datasets are introduced, each consisting of one of the following image modifications: screenshot simulation, social media filter simulation, meme filters, and Super-Resolution (SR). Specifically, the modifications are applied to the previously used test sets, which encompasses 638 images for the MJv6.1 dataset and 1000 images for all other datasets. This selection of modifications is made because they closely resemble real-life scenarios, which are not currently well-represented in state-of-the-art datasets for AID. The exact procedure for each of these modifications is described below.

Social media is widely used by everyone and is often the final destination for AI-generated images, as users share them to reach each other. Therefore, simulating a screenshot of a social media platform displaying an image closely resembles real-life scenarios. To simulate this, a template HTML page was designed to closely mimic such a social media platform. Using the Python library Selenium, images are dynamically injected into the HTML of the template page. To ensure each generated screenshot is unique, several dynamic elements were randomized on the page, including like count, comment count, share count, post status (saved or not), battery amount, current connection, account avatar, account name, image caption, and current time.

Regarding the social media filter simulation, it is decided to simulate Instagram filters due to its prevalence [3]. To simulate such filters, Pilgram2, a Python library consisting of Instagram-like filters [11], is used. For each image in a data folder, one of 40 filters is randomly applied to the image, as any filter can occur in real-world scenarios. To compress the images similar to Instagram's own image compression, we utilize JPEG compression with a compression-level of 75% on each image with a filter. The choice of the compression-level is motivated by the fact that Instagram uses a default compression-level between 70-80% [31].

The BSRGAN [32] is a super-resolution model and used to modify both fake and real images. Based on [4], applying super-resolution can camouflage artifacts introduced by image generators, but is only effective on images that are not fully synthetically generated. However, applying super-resolution to real images can be approached from two sides. The first is viewing the image as fake after the usage of super-resolution, since there are pixels that have been generated to create a larger image, which could create artifacts alongside. The second option is treating the image as real even after the super-resolution is used, as the understanding of the image is preserved. This is the perspective used in this work. The authors in [4] note that using super-resolution on real images can cause confusion by image detectors and degrade AID reliability. The BSRGAN is applied to both the synthetic images and real images from the original dataset. The original image is downscaled by a factor of K, after which super-resolution is applied to reconstruct an image of the same size as the original.

The final augmentation is an image meme filter. The idea behind this is similar to the screenshot simulation, but inverted. Rather than adding noise outside of the image, noise is added inside of the image. This is done through occlusion of the image. Two types of occlusion are used, inspired by meme images on the Internet. The first type consists of text at the top and bottom of the image. Since occlusion is the main focus, any text can be utilized and its content does not matter. To acquire a set of sentences, the first two chapters of the Bible were used. A sentence is randomly chosen and added to the image. Since there are only 2500 sentences and 10000 images that need to be created, noise is added to the text to avoid overfitting. There is a random chance for two letters in the text to be swapped. This creates more than 10,000 distinct sentences. The second type of occlusion is emojis that are pasted into the image. This once again occludes the image and distorts the spectral distribution. Other measures are also taken to avoid overfitting. The horizontal and vertical positions of the text are translated by a small randomized offset, the font type and font size are randomized. Furthermore, the emoji size and emoji amount are also randomized. All these randomizations are within a certain interval. They ensure that the model will likely not learn any aspect of the augmentations other than the original image.

4.2 Integrated Model Approach

To improve the robustness and generalization capability of AID, we propose a hybrid model (ROGER) that combines the following three state-of-the-art detector models [29]: SPAI [15], RINE [16], and PatchCraft [33]. Each model captures different aspects of the image, offering complementary perspectives. Specifically, RINE focuses on mid-level representations by extracting features from

intermediate layers of CLIP [16]. PatchCraft emphasizes low-level texture inconsistencies through patch-based analysis and inter-pixel correlations [33]. SPAI leverages low-level spectral distributions by learning to reconstruct masked frequency components. To combine these diverse representations, the final feature representations is extracted from each model, projected into a 256-dimensional space and normalized using a layer normalization. These three embeddings are concatenated into a unified feature vector. This 768-dimensional representation is then fed into a Multi-Layer Perceptron (MLP). This MLP has 2 layers with a size of 1536 with ReLU as activation function and dropout of 0.5 and a final layer to predict the probability that the image is fake. The MLP classifier is trained on 35,994 images (17,997 real, 17,997 generated, with each having 16,198 training images and 1799 validation images) sourced from the original SPAI training dataset [15]. This combined approach makes use of the strengths of each individual model, allowing for a detailed analysis of both spatial and frequency domain features, leading to improved detection performance.

4.3 Results

The evaluation of SPAI on the four modified datasets is displayed in Table 2. As seen in this Table, each of the datasets deteriorates the performance of SPAI, with the screenshot data being the most detrimental. This observation supports the notion that SPAI is not robust against real-world contexts.

Table 2: SPAI's evaluation on the modified datasets. The modified datasets identify the weaknesses of SPAI, as it decreases SPAI's performance during inference with respect to the baseline (SPAI on data without RAISE). The decrease in performance is displayed in red.

Image Size	<0.5 MPixels			0.5 - 1.0 MPixels						>1.0 MPixels				**AVG**
Approach	Glide	SD1.3	SD1.4	Flux	DALLE2	SD2	SDXL	SD3	GigaGAN	MJv5	MJv6.1	DALLE3	Firefly	
SPAI on data without RAISE	91.1	99.6	99.6	84.7	92.1	96.9	97.7	78.1	86.9	95.0	85.5	91.3	96.4	**91.9**
SPAI on meme data	63.1	94.3	95.1	74.5	81.9	86.3	92.1	65.5	81.0	90.0	78.1	78.8	74.8	**81.2** (10.7 ↓)
SPAI on SR data	77.7	94.6	95.1	72.1	71.4	87.9	91.4	38.5	69.8	89.9	79.8	92.4	88.0	**80.6** (11.3 ↓)
SPAI on Instagram filter data	76.4	87.6	88.5	70.0	74.7	74.0	79.3	66.9	66.3	68.2	68.7	76.4	68.3	**74.3** (17.6 ↓)
SPAI on screenshot simulation data	80.3	92.9	93.3	41.5	51.1	52.3	67.4	51.4	63.8	51.5	52.3	49.3	62.2	**62.3** (29.6 ↓)

When assessing ROGER on the same four datasets, we observe an inverse trend in performance. In Table 2, each evaluation of ROGER is compared to that of SPAI, with the respective increase in performance being noted. In all cases, ROGER outperforms SPAI in the average AUC, making it more robust than SPAI in those cases.

Table 3: ROGER (ours) evaluation on the modified datasets. The displayed differences are performance increases with respect to the corresponding results from Table 2.

Image Size	<0.5 MPixels			0.5 - 1.0 MPixels						>1.0 MPixels				**AVG**
Approach	Glide	SD1.3	SD1.4	Flux	DALLE2	SD2	SDXL	SD3	GigaGAN	MJv5	MJv6.1	DALLE3	Firefly	
ROGER on data without RAISE	98.9	100.0	100.0	98.9	99.1	100.0	100.0	67.0	99.9	99.7	99.7	84.0	98.1	**95.8** (3.9 ↑)
ROGER on meme data	94.4	99.8	99.8	95.9	97.2	99.6	99.9	51.7	99.5	98.9	98.8	62.2	79.8	**90.6** (9.4 ↑)
ROGER on SR data	99.8	100.0	100.0	99.9	99.9	100.0	100.0	46.3	99.9	100.0	100.0	100.0	100.0	**95.8** (15.2 ↑)
ROGER on Instagram filter data	82.4	91.3	91.8	62.4	83.1	86.3	79.5	55.6	68.7	71.9	60.5	64.3	85.0	**75.6** (1.3 ↑)
ROGER on screenshot simulation data	90.9	98.1	98.2	90.1	95.7	92.3	97.1	69.1	98.6	91.2	97.0	74.4	92.4	**91.2** (28.9 ↑)

5 Conclusion

In this paper, we introduced a pipeline consisting of four robustness tests designed to evaluate performance on derivative images. Furthermore, we introduced four new datasets that represent real-world scenarios for assessing the robustness of AID models. Additionally, we proposed the RObust AI-GEnerated image Recognizer (ROGER), which can generalize to both derivative and non-derivative images. We find that ROGER exhibits a higher performance when compared to SPAI and generalizes better to derivative images. However, since ROGER utilizes the embeddings of three models, it is slower and more computationally expensive than SPAI. The main limitation of ROGER is that it is entirely dependent on the three base models (SPAI, RINE, and PatchCraft). If all these models fail to perform well, the resulting performance will also be low. For future work, we could fine-tune both SPAI and ROGER on augmented and derivative images and whether they will improve in performance.

References

[1] Quentin Bammey. "Synthbuster: Towards detection of diffusion model generated images". In: *IEEE Open Journal of Signal Processing* 5 (2023), pp. 1–9.

[2] James Betker et al. "Improving image generation with better captions". In: *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf* 2.3 (2023), p. 8.

[3] Yu-Hsiu Chen et al. "Filter-invariant image classification on social media photos". In: *Proceedings of the 23rd ACM international conference on Multimedia*. 2015, pp. 855–858.

[4] Davide Alessandro Coccomini et al. *Exploring Strengths and Weaknesses of Super-Resolution Attack in Deepfake Detection*. 2024. arXiv: 2410.04205 [cs.CV]. URL: https://arxiv.org/abs/2410.04205.

[5] Riccardo Corvi et al. "On The Detection of Synthetic Images Generated by Diffusion Models". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10095167.

[6] Duc-Tien Dang-Nguyen et al. "Raise: A raw images dataset for digital image forensics". In: *Proceedings of the 6th ACM multimedia systems conference*. 2015, pp. 219–224.

[7] Jia Deng et al. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[8] Prafulla Dhariwal and Alex Nichol. *Diffusion Models Beat GANs on Image Synthesis*. 2021. arXiv: 2105.05233 [cs.LG]. URL: https://arxiv.org/abs/2105.05233.

[9] Mohamed Elasri et al. "Image Generation: A Review". In: *Neural Processing Letters* 54 (Mar. 2022). DOI: 10.1007/s11063-022-10777-x.

[10] Patrick Esser et al. "Scaling rectified flow transformers for high-resolution image synthesis". In: *Forty-first international conference on machine learning*. 2024.

[11] Jayshree Gupta, Sumeet Saurav, and Sanjay Singh. "Analyzing the Impact of Instagram Filters on Facial Expression Recognition Algorithms". In: *International Conference on Computer Vision and Image Processing*. Springer. 2023, pp. 152–163.

[12] Benjamin Hadwiger and Christian Riess. "The Forchheim image database for camera identification in the wild". In: *International Conference on Pattern Recognition*. Springer. 2021, pp. 500–515.

[13] Chih-Chung Hsu, Yi-Xiu Zhuang, and Chia-Yen Lee. "Deep Fake Image Detection Based on Pairwise Learning". In: *Applied Sciences* 10.1 (2020). ISSN: 2076-3417. DOI: 10.3390/app10010370. URL: https://www.mdpi.com/2076-3417/10/1/370.

[14] Minguk Kang et al. "Scaling up gans for text-to-image synthesis". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 10124–10134.

[15] Dimitrios Karageorgiou et al. "Any-Resolution AI-Generated Image Detection by Spectral Learning". In: *arXiv preprint arXiv:2411.19417* (2024).

[16] Christos Koutlis and Symeon Papadopoulos. "Leveraging Representations from Intermediate Encoder-Blocks for Synthetic Image Detection". In: *Computer Vision – ECCV 2024*. Ed. by Aleš Leonardis et al. Cham: Springer Nature Switzerland, 2025, pp. 394–411. ISBN: 978-3-031-73220-1.

[17] Alina Kuznetsova et al. "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale". In: *International journal of computer vision* 128.7 (2020), pp. 1956–1981.

[18] LatentSpacer. Civitai: Dataset with 6000+ FLUX.1 [dev] Images - 1024x768 and 768x1024. [Online]. Available from: https://civitai.com/models/631007?modelVersionId=705402. [Accessed April 30, 2025].

[19] Tsung-Yi Lin et al. "Microsoft coco: Common objects in context". In: *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*. Springer. 2014, pp. 740–755.

[20] Bo Liu et al. "Detecting Generated Images by Real Images". In: *Computer Vision – ECCV 2022*. Ed. by Shai Avidan et al. Cham: Springer Nature Switzerland, 2022, pp. 95–110. ISBN: 978-3-031-19781-9.

[21] Gustav Müller-Franzes et al. "A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis". In: *Scientific Reports* 13.1 (July 2023), p. 12098. ISSN: 2045-2322. DOI: 10.1038/s41598-023-39278-0. URL: https://doi.org/10.1038/s41598-023-39278-0.

[22] Alex Nichol et al. “Glide: Towards photorealistic image generation and editing with text-guided diffusion models”. In: *arXiv preprint arXiv:2112.10741* (2021).

[23] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. “Towards Universal Fake Image Detectors That Generalize Across Generative Models”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2023, pp. 24480–24489.

[24] Yogesh Patel et al. “Deepfake Generation and Detection: Case Study and Challenges”. In: *IEEE Access* 11 (2023), pp. 143296–143323. DOI: 10.1109/ACCESS.2023.3342107.

[25] Ryan Po et al. *State of the Art on Diffusion Models for Visual Computing*. 2023. arXiv: 2310.07204 [cs.AI]. URL: https://arxiv.org/abs/2310.07204.

[26] Dustin Podell et al. “Sdxl: Improving latent diffusion models for high-resolution image synthesis”. In: *arXiv preprint arXiv:2307.01952* (2023).

[27] Aditya Ramesh et al. “Hierarchical text-conditional image generation with clip latents”. In: *arXiv preprint arXiv:2204.06125* 1.2 (2022), p. 3.

[28] saq1b. Huggingface: saq1b/midjourney-v6.1. [Online]. Available from: https://huggingface.co/datasets/saq1b/midjourney-v6.1/tree/main. [Accessed April 30, 2025].

[29] Manos Schinas and Symeon Papadopoulos. “SIDBench: A Python framework for reliably assessing synthetic image detection methods”. In: *Proceedings of the 3rd ACM International Workshop on Multimedia AI against Disinformation*. 2024, pp. 55–64.

[30] Sandeep Singh Sengar et al. “Generative artificial intelligence: a systematic review and applications”. In: *Multimedia Tools and Applications* (Aug. 2024). ISSN: 1573-7721. DOI: 10.1007/s11042-024-20016-1. URL: https://doi.org/10.1007/s11042-024-20016-1.

[31] Slipsum Team. Why does Instagram ruin photo quality? [Online]. Available from: https://slipsum.com/why-does-instagram-ruin-photo-quality/. [Accessed May 11, 2025].

[32] Kai Zhang et al. *Designing a Practical Degradation Model for Deep Blind Image Super-Resolution*. 2021. arXiv: 2103.14006 [eess.IV]. URL: https://arxiv.org/abs/2103.14006.

[33] Nan Zhong et al. “Patchcraft: Exploring texture patch for efficient ai-generated image detection”. In: *arXiv preprint arXiv:2311.12397* (2023).