Team RoMa @ AADD-2025: On the Generation of Transferable and Visually Imperceptible Adversarial Attacks Against Deepfake Detectors

Nicolas Göller
Fraunhofer SIT / ATHENE
Darmstadt, Hesse, Germany
nicolas.goeller@sit.fraunhofer.de

Lukas Graner
Fraunhofer SIT / ATHENE
Darmstadt, Hesse, Germany
lukas.graner@sit.fraunhofer.de

Raphael Antonius Frick
Fraunhofer SIT / ATHENE / TU-Darmstadt
Darmstadt, Hesse, Germany
raphael.frick@sit.fraunhofer.de

Niklas Bunzel
Fraunhofer SIT / ATHENE / TU-Darmstadt
Darmstadt, Hesse, Germany
niklas.bunzel@sit.fraunhofer.de

Abstract

The rapid development of generative AI and in particular deepfake technology enables the seamless creation and manipulation of visual content. As the resulting syntheses are often indistinguishable from authentic images, they threaten the integrity of visual evidence. While forensic detectors can be used to detect syntheses, they can become targets of adversarial attacks. In the "Adversarial Attacks on Deepfake Detectors" challenge, competitors were tasked with perturbing a dataset of AI-synthesized images so that four classifiers would mistakenly accept them as authentic. In this paper, we introduce our solution, a white-box adversarial framework that injects globally distributed, data-driven noise perturbations optimized via additional surrogate Vision Transformer and EfficientNet classifiers. Empirical comparisons to both conventional post-processing transforms and localized adversarial patches demonstrate that our approach based on globally distributed noise achieves the highest attack success rates across all public detectors while preserving superior SSIM, confirming its efficacy and visual imperceptibility. In the final evaluation of the challenge, our proposed approach placed third with a final score of 2679.

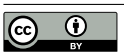
CCS Concepts

• **Security and privacy** → *Social network security and privacy*; • **Information systems** → *Multimedia information systems.*

Keywords

Adversarial Attacks, Deepfake Detection, Counter Forensics

ACM Reference Format:
Nicolas Göller, Lukas Graner, Raphael Antonius Frick, and Niklas Bunzel. 2025. Team RoMa @ AADD-2025: On the Generation of Transferable and Visually Imperceptible Adversarial Attacks Against Deepfake Detectors. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3746027.3761984



MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3761984

1 Introduction

Over the past decade, the amount of digital images and video circulating across Internet platforms has grown significantly. Such multimedia content is often featured in journalistic publications and disseminated via social media channels to substantiate narratives, to illustrate arguments, or to engage with audiences.

Concurrently, advances in artificial intelligence (AI) research and hereby most profoundly in generative adversarial networks (GANs) [9] and diffusion-based models [13], have dramatically increased the ability to synthesize entirely new images or to manipulate authentic images. As a result, AI-driven synthesis and manipulation of imagery now present a substantive threat to the integrity of visual evidence, driving an urgent need for forensic authentication techniques capable of discriminating genuine content from artificially generated or manipulated media.

Over the recent years, a variety of detection techniques have been developed that try to detect generated images. Although some achieve high accuracy on known benchmark-datasets, most struggle to generalize to new, unseen examples, i.e., they tend to overfit their training data, and suffer from performance loss under heavy post-processing or due to targeted adversarial attacks. To guard against these vulnerabilities, one direction is to craft a comprehensive benchmark dataset that researchers can use to develop and evaluate novel defense techniques.

In the "Adversarial Attacks on Deepfake Detectors: A Challenge in the Era of AI-Generated Media", participants received a collection of AI-created images and had to alter them so that a set of deepfake detectors misclassifies them as genuine. Every team has access to two detection models, while in the final testing round, two additional, held-out detectors are introduced to simulate black-box conditions and test an attack's cross-model generalization.

Each submitted method is scored based on the combination of two metrics: its misclassification rate against all four detectors (attack success rate) and the perceptual fidelity of the manipulated images, quantified via the Structural Similarity Index (SSIM) [17]. This ensures that solutions not only transfer effectively to unseen models but also remain as visually imperceptible as possible.

This paper presents our solution to solving the challenge. Under the premise that the target detectors are data-driven models, we

developed an adversarial framework that introduces noise perturbations into the AI-generated images. In our experiments against both post-processing filters and localized patch-based attacks, our noise-perturbation approach yielded higher perceptual fidelity while preserving equally strong or better attack success rates.

2 Adversarial Attacks on Deepfake Detectors: A Challenge in the Era of AI-Generated Media

The challenge was dedicated towards the development of adversarial attacks that can fool deepfake detection methods trained on AI-generated images derived from Generative Adversarial Networks (GANs) and Diffusion models.

2.1 Challenge Dataset

The challenge dataset consisted exclusively of deepfake images, all of which were synthesized using either GANs or diffusion-based models [1].

GANs consist of two neural networks [9]. A generator and a discriminator network compete in a zero-sum game in which the generator learns to map random vectors to realistic images, while the discriminator learns to distinguish them from real images. Training proceeds iteratively, with the generator improving its output to fool the discriminator and the discriminator improving to identify fakes.

Diffusion models, by contrast, begin by progressively degrading training images with noise and then learn a reverse diffusion process [13]. For this, a neural network is trained to predict the current image noise and to denoise them step by step, transforming pure noise into coherent images.

The primary data for this challenge were drawn from the WILD dataset [1], created for deepfake detection and model attribution. For the challenge a subset containing image splits featuring two resolutions were provided: a low-quality set (256×256 px) and a high-quality set (512×512 px - 1024×1024 px). In total, the following generative models produced 693 high-quality (HQ) and 710 low-quality (LQ) images: Adobe Firefly (Diffusion, HQ), Deep AI (Diffusion, HQ + LQ), Flux.1.1 Pro (Diffusion, HQ), Flux.1 (Diffusion, LQ), Freepik (Diffusion, LQ), Hotpot AI (Diffusion, HQ + LQ), Nvidia Sana PAG (Diffusion, HQ + LQ), Stable Diffusion 3.5 (Diffusion, HQ), Stable Diffusion Attend and Excite (Diffusion, LQ), StyleGAN2 (GAN, HQ + LQ), StyleGAN3 (GAN, HQ + LQ), Tencent Hunyuan (Diffusion, HQ + LQ).

2.2 Models

The challenge encompassed four target networks: a ResNet-50, a DenseNet-121, a ViT-B16, and a DenseNet-121-DCT model. During the development phase of the challenge, only the RGB-based ResNet-50 and DenseNet-121 architectures were provided to participants. In the evaluation phase, our adversarial examples were further assessed against a ViT-B16 model and another detector operating on DCT amplitude inputs (DenseNet-121-DCT).

[1] AADD-Challenge Code, Dataset & Leaderboard

2.3 Metrics

To quantify the effectiveness and imperceptibility of the developed adversarial attacks, the challenge employs two complementary metrics:

- **Attack Success Rate (ASR):** The proportion of adversarial examples that a detector misclassifies as "real":

$$\sum_{C_f \in C} C_f(I_k^{\text{ADV}}) = \text{LABEL}_{\text{real}}$$

  where C refers to a set of classifiers with $f \in (1 \dots N)$ and $N = 4$ as well as I_k^{ADV} being the k-th image that has been adversarially modified.
- **Structural Similarity Index (SSIM):** The SSIM is perceptual-quality metric [17] in $[0, 1]$ that quantifies the visual similarity between an original k-th deepfake image I_k and its perturbed version I_k^{ADV}:

$$\sum_{k=1}^{K} \text{SSIM}(I_k, I_k^{\text{ADV}})$$

  where K refers to the number of images I within the dataset.
- **Challenge Score:** The challenge score then combines both metrics, measuring the attack's success rate and imperceptibility:

$$\text{CS} = \sum_{C_f \in C} \sum_{k=1}^{K} \text{SSIM}(I_k, I_k^{\text{ADV}}) \cdot [C_f(I_k^{\text{ADV}}) = \text{LABEL}_{\text{real}}]$$

  where the brackets ([and]) will apply floor-based rounding. Higher values of CS indicate attacks that both evade detection consistently and remain visually indistinguishable.

3 Attacks on Deepfake Detectors

Deepfake detection methods can be broadly split into two categories. Model-driven techniques take advantage of handcrafted features often based on rules derived from signal-processing [8, 16]. By this, they rely on considerable expert knowledge. In contrast, data-driven detectors employ deep neural networks trained end-to-end on real and synthesized imagery [11], automatically learning the most discriminative features for classification. Although these learned models often perform well on benchmark datasets, they frequently overfit to the training distribution and hence do not generalize well towards unseen examples. To improve generalization and robustness against data-variation, modern pipelines integrate data-augmentation during training, ranging from geometric transforms (rotations, scaling, perspective warps) and the introduction of color variations (brightness, contrast, color jitter).

Deepfake detectors can be circumvented through three adversarial strategies: (1) conventional post-processing transformations applied to the entire image (e.g., blurring, compression, geometric warps), (2) globally distributed noise perturbations [5], and (3) localized adversarial patches inserted in specific regions of the images [19]. Each strategy comes with its distinct trade-offs in terms of efficacy and visual fidelity.

To analyze, which method are most suitable for solving the challenge, we performed an empirical evaluation study on the various solutions.

(a) Probability: 0.45% SSIM: 0.192 (b) Probability: 42% SSIM: 0.520 (c) Probability: 57% SSIM: -0.070 (d) Probability: 100% SSIM: 0.809 (e) Probability: 100% SSIM: 0.399

Figure 1: Examples of conventional post-processing applied to images with fake class probability (ResNet-50 classifier) & SSIM.

(a) Original image Probability: 100% (b) Attacked image Probability: 0.54% SSIM: 0.9384

Figure 2: Example of an effective patch attack applied to images with fake class probability (DenseNet-121 classifier).

(a) HQ image Prediction: authentic SSIM: 0.9183 (b) LQ image Prediction: authentic SSIM: 0.9974

Figure 3: Examples of globally distributed noise perturbations applied to images.

3.1 Conventional Post-Processing

In the past, a range of image-processing operations was found to mislead deepfake detectors successfully [2, 4]. For instance, deepfake detectors concerning face swapping detection are often prone to underperform when the material has been highly compressed. In addition, random noise, i.e., noise that has been applied for improved visuals, can also lead to a decrease in accuracy. To systematically study these effects, we leveraged the Kornia library [12] to apply a variety of post-processing transforms to our test images.

Examples of this evaluation can be seen in Figure 1. Only a few of the tested operations actually lowered the public classifiers' accuracy, implying that those models were probably trained with heavy data augmentation that builds robustness against most post-processing attacks. Moreover, the few manipulations that do succeed reduce the SSIM, making them impractical for this challenge. As a result, accuracy stays nearly constant when applying post-processing techniques while SSIM values decrease dramatically.

3.2 Localized Patch-Based Attacks

Localized patch attacks embed noise into certain regions of an image [19]. However, our tests revealed that in order to fool a classifier, they must cover fairly large areas, which inevitably harms perceptual fidelity and has a strong impact on the SSIM-score (Figure 2). Yet in cases where concealment is not the goal (for example, for patches attached on objects in real-world), they still serve as a straightforward way to mislead vision models.

3.3 Globally Distributed Noise Attacks

In contrast to the aforementioned strategies, the global noise-perturbation scheme consistently achieved the highest attack success rates across all known detectors during the development phase and maintained high SSIM scores (Figure 3), demonstrating the best balance between efficacy and visual imperceptibility.

4 Proposed Solution

In the following, we present a comprehensive description of our approach and report performance metrics obtained on the models in both phases. Since our empirical evaluation revealed that globally applied noise perturbations consistently outperform both post-processing and patch-based attack strategies, we decided to craft noise perturbations to fool the deepfake classifiers.

4.1 Setup & Implementation

Since the exact target models were unknown during the development phase of the challenge, we also optimized on surrogate networks that match popular deepfake detectors to enable transferability. Optimizing attacks for transferability against many models makes the noise stronger and lowers image quality (SSIM). Attacks tuned for specific models stay degrade the image less. As a result, we split our approach: for low-resolution images, we kept perturbations small to preserve quality, even if they transfer less; for high-resolution images, we allowed stronger perturbations, so the attacks would work more reliably on unseen models.

In particular, we followed the following steps:

Model	Attack Success	Mean SSIM (LQ)	Mean SSIM (HQ)	Succ. Attacks (LQ)	Succ. Attacks (HQ)
ResNet-50	0.9857	0.9967	0.8692	704	679
DenseNet-121	0.9800	0.9967	0.8692	697	678
ViT B-16	0.0321	0.9967	0.8692	5	40
DenseNet-121 DCT	0.0385	0.9967	0.8692	21	33

Table 1: Attack success rates, SSIM metrics, and number of successful attacks on low-quality (LQ) and high-quality (HQ) image subsets for various models.

(1) **Surrogate Model Training:** To enable white-box adversarial optimization despite unknown hold-out detectors, we trained two surrogate models on the challenge's generated images augmented with high-resolution face data from the FFHQ dataset. The FFHQ dataset was originally used in the training of StyleGAN and its derivates, making it a suitable candidate. Based on it, we implemented a Vision Transformer (ViT-B16) [3] and an EfficientNet-B0 [14] as additional classifiers. These architectures were selected because they can often be found in state-of-the-art deepfake detectors. Moreover, EfficientNet-B0 features resource-efficient convolutional feature learning and ViT-B16 provides global self-attention mechanism [15], which captures image dependencies in a manner orthogonal to traditional CNNs like the provided ResNet [6] or DenseNet [7].

(2) **Noise Perturbation Generation:** For the noise generator dealing with low-quality images, we first initialize the perturbation tensor to zero. Next, we project it into the image by multiplying it element-wise with the target's RGB channels. The perturbed image is then passed through all four classifiers (two public, two surrogate), producing a composite score that combines attack success rate with SSIM. Finally, we back-propagate this score to iteratively refine the noise until we achieve our desired balance of misclassification and visual fidelity. We employ the Adam optimizer [10] to update the noise perturbations, taking advantage of its adaptive learning rates and momentum terms to ensure stable, fast convergence. For the initial learning rate, we took advantage of a learning rate of 0.0005. As the image resolution for the images of the low-quality split was fixed to a resolution of 256×256 px, the images were processed in batches of 10.

    For the high-quality images, we deployed DI-FGSM (Diverse-Input Iterative Fast Gradient Sign Method) [18]. It is an extension of the classic FGSM family [5] that injects stochastic transformations into each gradient-based update. At each of our 10 iterations, the input is randomly resized and padded, then a small FGSM step is applied to the transformed image. Repeating this over multiple rounds yielded perturbations that remained effective even when transferred to unknown, surrogate black-box models. Since the input images span arbitrary resolutions, we run the optimization loop on each image individually rather than in mini-batches for a maximum of 1000 iterations. This per-sample strategy accommodates variable dimensions without requiring forced resizing or padding for batch uniformity.

(3) **Dataset Finalization:** The optimized noise perturbations are applied onto the images and saved to disk.

4.2 Results

Using the evaluation script provided by the organizers of the challenge, an estimation of the overall challenge score as well as the resulting SSIM and attack success rate for each of the public classifiers could be obtained. The method achieved an attack success rate of 0.9857 on the ResNet detector and 0.9800 on the DenseNet model (Table 1). With a mean Structural Similarity Index (SSIM) of 0.9338, the final computed challenge score was 2589.

When assessed against the unseen classifiers from the evaluation phase, our score rose only slightly to 2679, while the SSIM remained unchanged at 0.9338 since the input images were identical. The minimal gain despite adding more models indicates that our perturbations did not generalize well to the new models. A closer inspection of each classifier's results (Table 1) confirms this finding. Only 45 of 1403 images classified with the ViT were incorrectly classified, whereas 54 of 1403 images were falsely classified as authentic by the DenseNet-121 DCT model. While both noise generators exhibited limited transferability, the DI-FGSM-based method achieved superior generalizability.

Although we included a ViT surrogate and optimized noise for a DenseNet model, differences in training likely impeded transfer. The undisclosed detectors were trained not only on FFHQ but also on Celeb-A as genuine data, and the one DenseNet variant analyzed DCT amplitudes rather than RGB values, factors that probably contributed to a reduced cross-model robustness.

5 Conclusion

In this paper, we describe our submission to the "Adversarial Attacks on Deepfake Detectors" challenge. Through empirical evaluation, we found that globally applied noise perturbations consistently outperformed alternative attack strategies. Accordingly, we adopted an iterative white-box optimization routine to craft these perturbations. Because only two of the four evaluation models were disclosed during development, we trained surrogate EfficientNet-B0 and Vision Transformer (ViT-B16) classifiers to approximate the remaining black-box detectors, and took advantage of a combination of noise generators, each with their own benefits and disadvantages. Validation on the two public classifiers shows that our noise-based attack achieves a challenge score of 2589, demonstrating both high attack success rates and low perceptual distortion. In the final evaluation, the approach placed third with a final score of 2679. A detailed breakdown, however, exposed that, even matching surrogate and target architectures, the crafted perturbations transferred poorly to the non-public models. This gap shows that true cross-model transferability remains unresolved and warrants further research as part of future work.

Acknowledgements

This research work has been funded by BMBF and the Hessen State Ministry for Higher Education, Research and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

References

[1] Pietro Bongini, Sara Mandelli, Andrea Montibeller, Mirko Casu, Orazio Pontorno, Claudio Vittorio Ragaglia, Luca Zanchetta, Mattia Aquilina, Taiba Majid Wani, Luca Guarnera, Benedetta Tondi, Giulia Boato, Paolo Bestagini, Irene Amerini, Francesco De Natale, Sebastiano Battiato, and Mauro Barni. 2025. WILD: a new in-the-Wild Image Linkage Dataset for synthetic image attribution. arXiv:2504.19595 [cs.MM] https://arxiv.org/abs/2504.19595

[2] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton-Ferrer. 2020. The DeepFake Detection Challenge Dataset. *ArXiv* abs/2006.07397 (2020). https://api.semanticscholar.org/CorpusID:219687616

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929 [cs.CV] https://arxiv.org/abs/2010.11929

[4] Raphael Antonius Frick and Martin Steinebach. 2024. One Detector to Rule Them All? On the Robustness and Generalizability of Current State-of-the-Art Deepfake Detection Methods. *Electronic Imaging* 36 (2024), 1–6.

[5] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. arXiv:1412.6572 [stat.ML] https://arxiv.org/abs/1412.6572

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs.CV] https://arxiv.org/abs/1512.03385

[7] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2018. Densely Connected Convolutional Networks. arXiv:1608.06993 [cs.CV] https://arxiv.org/abs/1608.06993

[8] Yonghyun Jeong, Doyeon Kim, Seungjai Min, Seongho Joe, Youngjune Gwon, and Jongwon Choi. 2021. BiHPF: Bilateral High-Pass Filters for Robust Deepfake Detection. arXiv:2109.00911 [cs.CV] https://arxiv.org/abs/2109.00911

[9] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and Improving the Image Quality of StyleGAN. arXiv:1912.04958 [cs.CV] https://arxiv.org/abs/1912.04958

[10] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs.LG] https://arxiv.org/abs/1412.6980

[11] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. 2024. Towards Universal Fake Image Detectors that Generalize Across Generative Models. arXiv:2302.10174 [cs.CV] https://arxiv.org/abs/2302.10174

[12] E. Riba, D. Mishkin, J. Shi, D. Ponsa, F. Moreno-Noguer, and G. Bradski. 2020. A survey on Kornia: an Open Source Differentiable Computer Vision Library for PyTorch. arXiv:2009.10521 [cs.CV] https://arxiv.org/abs/2009.10521

[13] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752 [cs.CV] https://arxiv.org/abs/2112.10752

[14] Mingxing Tan and Quoc V. Le. 2020. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. arXiv:1905.11946 [cs.LG] https://arxiv.org/abs/1905.11946

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention Is All You Need. arXiv:1706.03762 [cs.CL] https://arxiv.org/abs/1706.03762

[16] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. 2023. DIRE for Diffusion-Generated Image Detection. arXiv:2303.09295 [cs.CV] https://arxiv.org/abs/2303.09295

[17] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612. https://doi.org/10.1109/TIP.2003.819861

[18] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan Yuille. 2019. Improving Transferability of Adversarial Examples with Input Diversity. arXiv:1803.06978 [cs.CV] https://arxiv.org/abs/1803.06978

[19] Chenglin Yang, Adam Kortylewski, Cihang Xie, Yinzhi Cao, and Alan Yuille. 2020. PatchAttack: A Black-box Texture-based Attack with Reinforcement Learning. arXiv:2004.05682 [cs.CV] https://arxiv.org/abs/2004.05682