Rethinking Perturbation-Based Training-Free Method for Deepfake Face Detection

Houying Zhou†, Yanjun Deng‡, Hao Wang‡⋆
† Hangzhou Dianzi University Information Engineering College, Hangzhou, China
‡ School of Cyberspace Security, Hangzhou Dianzi University, Hangzhou, China

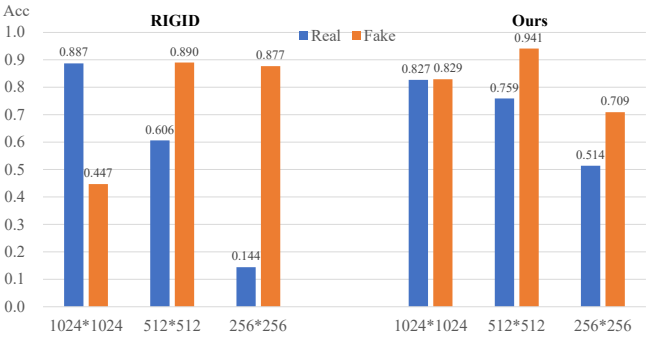
***Abstract*—Training-free deepfake image detection aims to discern whether inputs are authentic or synthetic by directly conduct evaluation on testing samples. Prior approaches predominantly measure the similarity between original images and perturbation-generated versions. Leveraging pre-trained foundation models like DINOv2, these methods typically deliver remarkable detection performance on natural images. However, their effectiveness diminishes significantly when applied to deepfake face detection, particularly for faces of varying resolutions. Additionally, the computational demands are high due to the complexity of these foundational models, hindering the application in real-world scenarios. To overcome these challenges, we elaborate a simple yet effective Upsampling-Perturbation-Downsampling method for training-free deepfake face detection. This approach enhances both the robustness against diverse input resolutions and the efficiency of detection process. Extensive experiments on our augmented DeepFakeFaceForensics dataset demonstrate that our approach significantly outperforms state-of-the-art methods.**

***Index Terms*—deepfake face detection, perturbation-based, training-free, resolution-robust, efficient detection**

I. Introduction

In recent years, deep generative methods, such as Generative Adversarial Networks (GANs) [1] and Diffusion-based Models (DMs) [2], have attracted ever-increasing attention in computer vision community and created numerous realistic-looking synthetic images. However, utilizing these generated content, especially fine-grained human facial images, results in the severe risks about misuse and poses a significant challenge for security in many fields. To address this issue, a variety of detection methods have been proposed to distinguish whether the facial image is authentic or synthetic.

For deepfake face detection, researchers commonly develop a two-class (i.e., real or fake) detector via mining forgery artifacts contained in the content. Early methods [3] [4] [5] discover that synthetic cues are generated through an upsampling operation, which is a crucial component of GANs or DMs for achieving resolution expansion. However, these specific artifacts are hard to be found with the advancement of generative methods. Hence, data-driven training-based detection approaches are proposed to mine common forgery cues across various generative models. These approaches can be roughly categorized into two types, i.e., spatial-based detection and frequency-based one. The former extracts local [6] [7] [8] [9] [10] or global [11] [12] [13] spatial features for detection. The latter mainly captures artifacts in frequency domain via Fast Fourier Transform (FFT) [14] analysis. Obviously, local-global spatial feature fusion [15] or spatial-frequency feature fusion [16] can further improves detection performance by sufficiently exploring complementary information. Recently, with the development of vision-language models [17], researchers [18] [19] have introduced them into the field of deepfake image detection and achieved impressive performance. Unfortunately, these training-based methods consume large amount of computational resource, which are extremely unfavorable for the large-scale training dataset. To alleviate this problem, AEROBLADE [20] conducts direct evaluation on testing samples through computing the reconstruction error of a pre-trained auto-encoder from DMs, opening up new avenues for training-free deepfake image detection. Furthermore, RIGID [21] utilizes perturbation-based method via measuring the similarity between original inputs and perturbation-generated version, achieving state-of-the-art detection performance.



Augmented Unprocessed Subset	RIGID	Ours
Overall Accuracy	0.600	0.720 (+**0.120**)
AUC	0.625	0.798 (+**0.173**)
Inference Speed	248 ms/img	26.7 ms/img (~**10x Faster**)

Fig. 1. Existing perturbation-based training-free method, i.e., RIGID, tends to identify high-resolution images as authentic and recognize low-resolution ones as synthetic. Our approach significantly enhances both the robustness against diverse input resolutions and the efficiency of detection process. It should be noted that the results are reported on augmented unprocessed subset of DeepFakeFaceForensics.

As illustrated in Fig. 1, although perturbation-based training-free methods like RIGID demonstrate promising results on detecting natural images, their effectiveness dimin-

⋆This work was supported by the Fundamental Research Funds for the Provincial Universities of Zhejiang (Grant GK249909299001-030). Hao Wang is the corresponding author (hwang@hdu.edu.cn).

ishes significantly when applied to deepfake facial image detection, particularly for faces of varying resolutions. Specifically, the images with higher resolution are prone to be recognized as genuine. Besides, they usually need a large amount of computational resources due to the complexity of foundation models (e.g., DINOv2 [22]), which play a role as feature extractor for subsequent similarity calculation. For example, RIGID averagely takes 268 millisecond to process one image on augmented unprocessed subset of DeepFakeFaceForensics [15], hindering the application in real-world scenarios. Although DINOv2 needs a large amount of computational resources, it performs much better than CLIP [17] or other state-of-the-art architectures.

In this paper, we propose a simple yet effective perturbation-based method to simultaneously circumvent these issues. Concretely, we elaborate a novel strategy, i.e., Upsampling-Perturbation-Downsampling (UPD), for training-free deepfake face detection, remarkably improving both robustness and efficiency. The main contributions of this work are as follows:

- To the best of our knowledge, it is the first work to explore perturbation-based training-free method for deepfake face detection, opening up new avenues for this task;
- We devise a novel method, Upsampling-Perturbation-Downsampling, to simultaneously enhance the robustness against diverse resolutions and detection efficiency;
- Experimental results on challenging DeepFakeFaceForensics dataset show that our method significantly outperforms state-of-the-art approaches.

II. Related Work

A. Training-Based Deepfake Detection

Traditional deepfake detection methods commonly train deep neural networks, such as Convolutional Neural Network (CNN) [23] and Vision Transformer (ViT) [24], to extract features for solving the binary (i.e., real or fake) classification task. They can be roughly categorized into two groups according the final features used for classification, i.e., spatial-based and frequency-based methods. The former mine local [6] [7] [8] [9] [10] or global [11] [12] [13] artifacts solely for detection. Specifically, Chai *et al.* [6] utilize patch-based classifier to mine the more easily detectable regions and achieve good generalization ability. Similarly, researchers [7] extract forgery cues from facial parts for detection. Multiple-attention network [8] adopts spatial attention mechanism to focus on different local parts. Patch-DFD [9] extracts five key patches around face based on the prior knowledge. To improve detection robustness, Mandelli *et al.* [10] adopt multiple orthogonal networks to calculate patch-level scores and then aggregate them for final classification. Meanwhile, global spatial features are also beneficial for detection. For example, CNN-Aug [11] achieves impressive generalization ability via extract global spatial features with the help of strong data augmentation. Recently, researchers introduce vision-language models into deepfake detection and show its advantage in extracting global spatial features for detection. UnivFD [18] utilizes fixed visual feature space of CLIP [17] to achieve universal fake image detection. Complementary to spatial features, frequency ones characterize forgery artifacts from different aspect. Researchers [14] propose frequency-based masking strategy and achieve slightly better results on universal fake image detection. Furthermore, conducting local-global spatial feature fusion [15] or spatial-frequency feature fusion [16] [19] significantly improves the detection performance. However, all these methods need training or fine-tuning on training samples, consuming much computational resources to obtain satisfactory performance.

TABLE I
The comparison between original unprocessed subset and augmented version of DeepFakeFaceForensics dataset.

(Resolution, Label)	Unprocessed Subset	Augmented Version
(256*256, Real)	-	3,160
(256*256, Fake)	3,160	3,160
(512*512, Real)	-	1,000
(512*512, Fake)	1,000	1,000
(1024*1024, Real)	1,000	2,000
(1024*1024, Fake)	2,000	2,000
Total	7,160	12,320

B. Training-Free Deepfake Detection

To alleviate the resource consumption, training-free deepfake detection methods emerge in recent years. UnivFD [18] utilizes nearest neighbor to determine whether the testing sample is authentic or synthetic. However, the training samples with annotations are needed for conducting nearest neighbor search. AEROBLADE [20] measures the similarity between original input and the reconstructed one, which is computed by a pre-trained auto-encoder of DMs. Obviously, AEROBLADE merely shows satisfactory detection performance on images synthesized by DMs. Hence, RIGID [21] proposes a perturbation-based method to measure the similarity between original input and transformed version, with the help of foundation models. It achieves impressive detection performance on natural photos while performs poor on detecting facial images. Besides, it still consumes much computational resources due to the high complexity of foundation models.

III. Methodology

A. Motivation

Firstly, we show the resolution bias of existing perturbation-based training-free methods, e.g., RIGID [21], through augmenting DeepFakeFaceForensics. Concretely, as demonstrated in Table I, the original dataset merely consists of authentic images with high resolution, i.e., 1024*1024, ignoring the evaluation on lower ones, such as 512*512 and 256*256. When simply conduct augmentation with downsampling operation, existing method performs poor on these augmented images, e.g., the Acc is only 0.144 on authentic samples of 256*256 resolution as illustrated in Fig. 1. This observation significantly weakens the effectiveness of existing perturbation-based training-free methods on deepfake face detection.

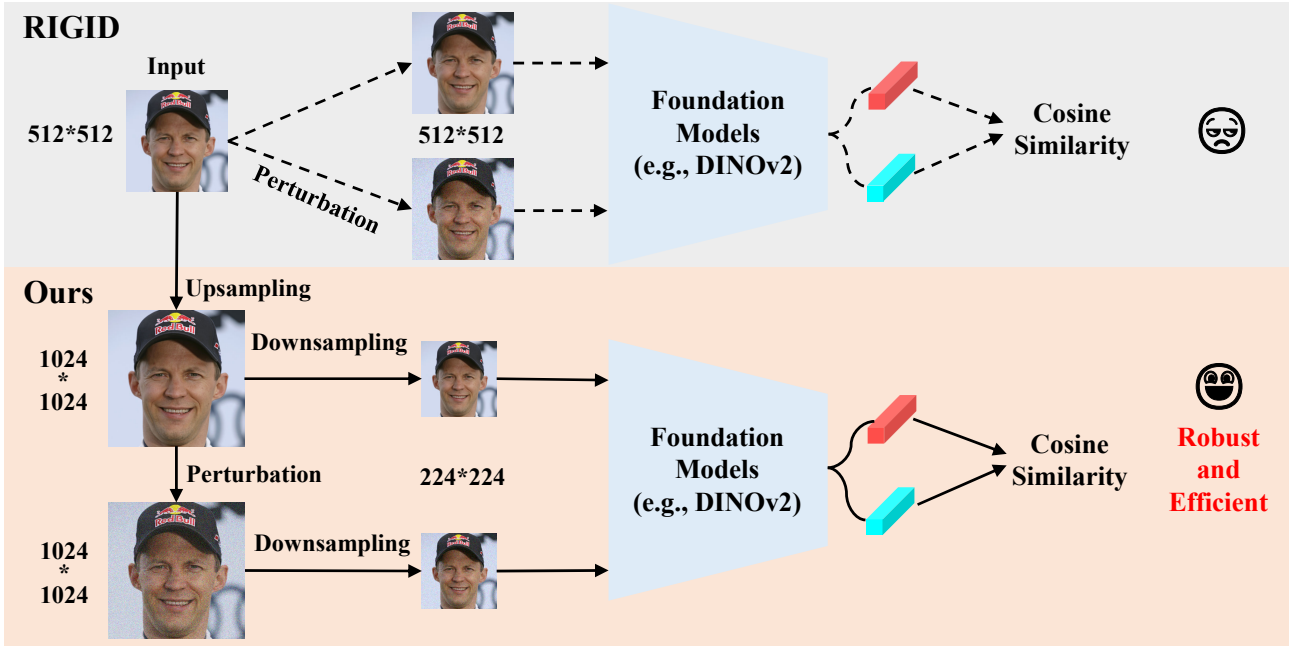


Fig. 2. The overall framework, which consists of upsampling, perturbation and downsampling operations.

Secondly, we attempt to answer "Why existing perturbation-based training-free methods fails to identify images of various resolution?". As is known to all, the core idea of perturbation-based training-free methods is measuring the distance between original input and perturbated one. This distance is not only affected by the intensity of injected noise used for perturbation but is also influenced by the smoothness of input content. Obviously, the latter decreases with a decline in resolution, making low-resolution samples more sensitive to be perturbated and then resulting in poor detection performance.

B. Upsampling-Perturbation-Downsampling

Hence, we propose a simple yet effective Upsampling-Perturbation-Downsampling (UPD) method to alleviate the issue of bias, as illustrated in Fig. 2. For low-resolution samples, we first conduct upsampling to obtain the version of high resolution, avoiding the low smoothness of input content. Formally, this process can be defined as follows:

$$x^{Up} = \begin{cases} x, & R(x) \geq R_{thr}, \\ Up(x), & R(x) < R_{thr}, \end{cases} \tag{1}$$

where x^{Up} and Up denote the upsampled version of input x and Upsampling function, respectively. $R(x)$ means the resolution of input x. R_{thr} represents the threshold to trigger the upsampling operation.

Then we follow the traditional perturbation method [21], i.e., injecting the noise drawn from a standard normal distribution $N(0, I)$, to obtain the perturbated samples. Specifically, the perturbation can be describe as follows:

$$x^{UpPert} = x^{Up} + \lambda * \delta, \quad \delta \sim N(0, I), \tag{2}$$

where x^{UpPert} is the perturbated version of upsampled input. δ denotes additive noise and λ controls its intensity.

After obtaining upsampled and perturbated inputs, we conduct downsampling to reduce the burden of computation for efficient detection, which can be formulated as follows:

$$x^{UpDown} = Down(x^{Up}), \tag{3}$$

$$x^{UpPertDown} = Down(x^{UpPert}), \tag{4}$$

where x^{UpDown} and $x^{UpPertDown}$ represent the downsampled version of upsampled and perturbated inputs. $Down$ denotes the downsampling operation.

C. Deepfake Detection

To measure the similarity between original input and transformed one, we utilize DINOv2 as backbone to extract features for subsequent similarity calculation. The decision of input image can be defined as follows:

$$\mathbb{I}(cos(F(x^{UpDown}),\ F(x^{UpPertDown})) \geq S_{thr}), \tag{5}$$

where F stands for feature extractor(e.g., ViT-L/14 [22]), cos represents the cosine similarity function, $\mathbb{I}$ denotes the binary indicator, and S_{thr} is the threshold to identify whether the input is authentic or synthetic.

IV. Experiment

A. Dataset and Experiment Setup

DeepFakeFaceForensics is a challenging dataset for deepfake face detction in real-world scenarios, which takes common post-processing (e.g., compression, blurring, adversarial sample, manipulation, etc.) and advanced generative models (e.g., GAN-based, DM-based, ViT-based, etc.) into consideration. It comprises of 6 subsets, i.e., unprocessed, common

TABLE II
EVALUATION RESULTS (AUC) ON AUGMENTED DEEPFAKEFACEFORENSICS. NOTED THAT GLFF AND RIGID ARE REPRODUCED.

Methods		Training Cost	Unprocessed	Post-processing	Anti-forensics	Multi-compression	Mixed	Average
Training-Based	**GLFF [15]**	∼16 GPU Days	0.677	**0.843**	**0.605**	0.360	**0.657**	0.628
Training-Free	**RIGID [21]**	0 GPU Days	0.625	0.661	0.447	0.542	0.463	0.547
	UPD (Ours)	0 GPU Days	**0.798**	0.835	0.499	**0.682**	0.477	**0.658**

TABLE III
EVALUATION RESULTS (OA) ON AUGMENTED DEEPFAKEFACEFORENSICS.

Test Data	GLFF [15]	RIGID [21]	UPD (Ours)
Unprocessed	0.599	0.600	**0.720**
Post-processing	0.546	0.604	**0.742**
Anti-forensics	**0.547**	0.392	0.495
Multi-compression	0.491	0.542	**0.629**
Mixed	**0.499**	0.473	0.475
Average	0.536	0.522	**0.612**

post-processing, face blending, anti-forensics and multi-image compression, with a total of 46,4000 synthetic images. For evaluation, researchers [15] randomly select 1,000 authentic images from FFHQ dataset [25].

Augmented DeepFakeFaceForensics is our re-constructed dataset via augmentation on original authentic images (i.e., 1024*1024 resolution) with downsampling, thus generating low-resolution samples, such as 512*512 and 256*256. Besides, we select more authentic images from FFHQ to make the number of real and fake images be equal. However, for face blending subset where the resolution of samples is usually more than 2560*2560, the computational burden (i.e., minutes per image) is unbearable for RIGID so that this subset is not augmented. We believe this augmented and balanced dataset can fairly evaluate the robustness of detection methods.

Evaluation Metrics includes Overall Accuracy (OA) and Area Under ROC Curve (AUC). It should be noted that accuracy and AUC are measured in each subset entirely, instead of computing the average across all generative models in each subset, like GLFF [15].

Following RIGID [21], we adopt ViT-L/14 [22] as feature extractor and set threshold S_{thr} as 0.95. The noise intensity λ is 0.08 and threshold R_{thr} is 1024. All experiments are conducted with PyTorch [26] on single RTX 4090 GPU.

B. Comparison with SOTA Methods

As illustrated in Table II and Table III, we reproduce and demonstrate detection results with 1 state-of-the-art training-based method, i.e., GLFF [15] and 1 training-free one, i.e., RIGID [21]. Firstly, our proposed approach achieves the best detection performance on both OA and AUC metrics, even surpassing the training-based method. Secondly, we achieve top 2 scores and 1 second-place score out of 5 subsets. For anti-forensics and mixed subsets, GLFF obtains the highest AUC and OA scores while RIGID and ours perform consistently, indicating that training-free methods have limitations in detecting these challenging deepfake facial images.

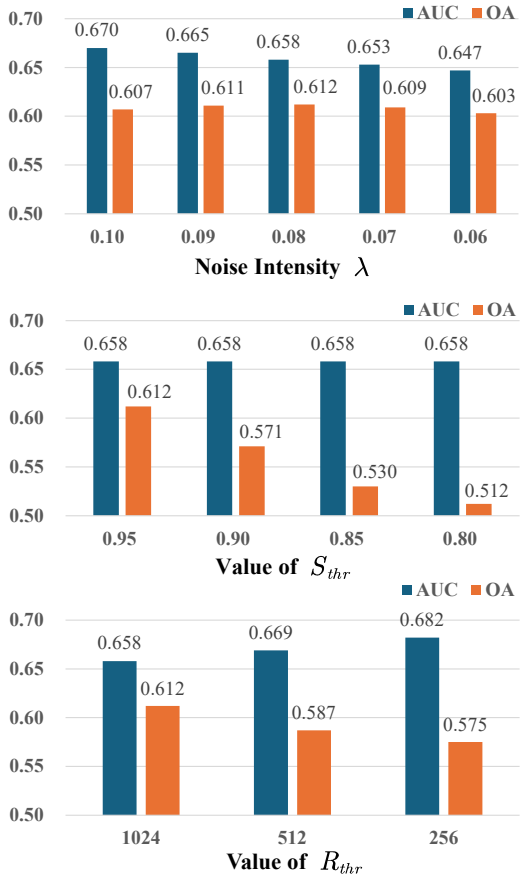


Fig. 3. Detection performance with different value of each hyper-parameter.

C. Ablation Studies

We conduct detail experiments with different value of each hyper-parameter and show the detection results (i.e., both AUC and OA) in Fig. 3. Specifically, our approach is robust with the change of noise intensity λ. However, the performance varies greatly with changes in the value of threshold S_{thr} and R_{thr}. To obtain better scores on both AUC and OA, we set the value of noise intensity λ, threshold S_{thr} and R_{thr} as 0.08, 0.95 and 1024, respectively.

V. CONCLUSION

In this paper, we have proposed a novel perturbation-based method, i.e., Upsampling-Perturbation-Downsampling, for deepfake face detection. Our approach simultaneously enhance the robustness and efficiency of training-free method. In the future, we should devote more efforts on anti-forensics and mixed subsets to improve the detection performance.

REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[2] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Neural Information Processing Systems*, vol. 33, 2020, pp. 6840–6851.

[3] K. Chandrasegaran, N.-T. Tran, and N.-M. Cheung, "A closer look at fourier spectrum discrepancies for cnn-generated images detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7200–7209.

[4] T. Dzanic, K. Shah, and F. Witherden, "Fourier spectrum discrepancies in deep network generated images," in *Neural Information Processing Systems*, 2020, pp. 3022–3032.

[5] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, "Leveraging frequency analysis for deep fake image recognition," in *International Conference on Machine Learning*, 2020, pp. 3247–3258.

[6] L. Chai, D. Bau, S.-N. Lim, and P. Isola, "What makes fake images detectable? understanding properties that generalize," in *European Conference on Computer Vision*, 2020, pp. 103–120.

[7] S. Schwarcz and R. Chellappa, "Finding facial forgery artifacts with parts-based detectors," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 933–942.

[8] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, "Multi-attentional deepfake detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2185–2194.

[9] M. Yu, S. Ju, J. Zhang, S. Li, J. Lei, and X. Li, "Patch-dfd: Patch-based end-to-end deepfake discriminator," *Neurocomputing*, vol. 501, pp. 583–595, 2022.

[10] S. Mandelli, N. Bonettini, P. Bestagini, and S. Tubaro, "Detecting gan-generated images by orthogonal training of multiple cnns," in *IEEE International Conference on Image Processing*, 2022, pp. 3091–3095.

[11] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "Cnn-generated images are surprisingly easy to spot... for now," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8695–8704.

[12] N. Yu, L. S. Davis, and M. Fritz, "Attributing fake images to gans: Learning and analyzing gan fingerprints," in *IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7556–7566.

[13] F. Marra, D. Gragnaniello, D. Cozzolino, and L. Verdoliva, "Detection of gan-generated fake images over social networks," in *IEEE Conference on Multimedia Information Processing and Retrieval*, 2018, pp. 384–389.

[14] C. T. Doloriel and N.-M. Cheung, "Frequency masking for universal deepfake detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024, pp. 13 466–13 470.

[15] Y. Ju, S. Jia, J. Cai, H. Guan, and S. Lyu, "Glff: Global and local feature fusion for ai-synthesized image detection," *IEEE Transactions on Multimedia*, vol. 26, pp. 4073–4085, 2023.

[16] C. Tian, Z. Luo, G. Shi, and S. Li, "Frequency-aware attentional feature fusion for deepfake detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

[17] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, 2021, pp. 8748–8763.

[18] U. Ojha, Y. Li, and Y. J. Lee, "Towards universal fake image detectors that generalize across generative models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 24 480–24 489.

[19] H. Liu, Z. Tan, C. Tan, Y. Wei, J. Wang, and Y. Zhao, "Forgery-aware adaptive transformer for generalizable synthetic image detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 10 770–10 780.

[20] J. Ricker, D. Lukovnikov, and A. Fischer, "Aeroblade: Training-free detection of latent diffusion images using autoencoder reconstruction error," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9130–9140.

[21] Z. He, P.-Y. Chen, and T.-Y. Ho, "Rigid: A training-free and model-agnostic framework for robust ai-generated image detection," *arXiv preprint arXiv:2405.20112*, 2024.

[22] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[24] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[25] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410.

[26] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Neural Information Processing Systems*, 2019.