

StealthDiffusion: Towards Evading Diffusion Forensic Detection through Diffusion Model

Supplementary materials

Anonymous Authors

In this supplementary material, we provide additional experimental and qualitative results to support the contributions in the main paper. In Section 1, we provide detailed information about the forensic techniques used to evaluate our attack method, followed by additional attack experimental results in Section 2. In Section 3, we demonstrate further image quality comparisons with baseline attacks.

1 FORENSIC DETECTION METHODS

In this section, we provide quantitative results of the forensic detection model used in the main paper to demonstrate that the detection method is able to successfully distinguish the differences between genuine and generated images. We trained four different backbones using the method proposed by [19] as our evidence detection model and adversarial model: Efficientnet-B0 [17], ResNet50 [8], DeiT [18] and Swin-T [9]. We further evaluated the black-box transfer attack performance of the latest detection methods, namely DIRE [20], GFF [10], Lgrad [16], RECCE [2] and UniDetection [14]. We fine-tuned these methods on the training set of GenImage [23] and presented their performance on the validation set GenImage in Tab. 1. All the detection methods maintained high Accuracy (ACC) and Area Under the Curve (AUC). By attacking high-performing detection models, we demonstrate the effectiveness of our method relative to other baseline approaches.

Table 1: The Accuracy(ACC) and Area Under the Curve(AUC) of the detection methods used in the main paper.

	ACC	AUC
Efficientnet-B0	97.58	99.42
ResNet50	95.56	99.09
DeiT	98.26	98.45
Swin-T	97.26	99.45
DIRE	99.99	99.99
GFF	99.73	99.89
Lgrad	97.66	99.73
RECCE	99.89	99.99
UniDetection	99.99	99.99

Permission to make digital or hard copies of all or part of this work for personal or internal use, or for the internal or personal use of specific clients, is granted by ACM for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
MM '24, 28 October - 1 November 2024, Melbourne, Australia
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/18/06
<https://doi.org/XXXXXXX.XXXXXXX>

2 ADDITIONAL EXPERIMENTAL RESULTS

Additional Core Component Analysis. Here, we generate attacks using Efficientnet-B0 (E) [17], ResNet50 (R) [8], and DeiT (D) [18], Tab. 2 presents the results of different variants of our method. For the baseline method, the Control-VAE and Latent Adversarial Optimization(LAO) methods are not used. Using either Control-VAE or LAO methods yields a positive effect on the ASR metrics for both backbones, and combining the two modules can bring more performance gains.

Table 2: Ablation study for core components of our method. The vertical E [17], R [8], D [18] are used to generate adversarial samples, The horizontal E [17], R [8], D [18], S [9] are used to detect adversarial samples.

	CVAE	LAO	E	R	D	S
E	✗	✗	100.00	60.50	47.38	75.13
	✓	✗	100.00	71.38	51.13	79.25
	✗	✓	100.00	79.25	53.75	81.50
	✓	✓	100.00	89.50	66.38	84.13
R	✗	✗	52.25	100.00	56.13	94.68
	✓	✗	52.50	100.00	58.38	95.00
	✗	✓	94.13	100.00	64.00	95.13
	✓	✓	94.75	100.00	67.25	95.25
D	✗	✗	78.75	64.00	100.00	96.50
	✓	✗	85.25	67.75	100.00	97.00
	✗	✓	88.88	89.13	100.00	97.38
	✓	✓	90.75	89.88	100.00	97.88

Additional Diffusion Steps and Optimizing iterations Analysis. We generate attacks using Efficientnet-B0 (E) [17], ResNet50 (R) [8], DeiT (D) [18] and Swin-T(S) [9]. In Tab. 3, we demonstrate the impact of Diffusion Steps and Optimizing iterations in the Latent Adversarial Optimization(LAO) process. A suitable value for Optimizing iterations can maximize the ASR of the method. After reaching a certain value, the impact on the success rate becomes less significant. With a smaller value for Optimizing Iterations, increasing Diffusion Steps will lead to a decrease in the method's ASR. This may be because the process introduces features of AI-generated images that the forensic classifier can recognize. With the increase of Optimizing Iterations, the method can still maintain a relatively high ASR. Considering factors such as image quality and optimization time, our method selects a value of 2 for Diffusion Steps and 5 for Optimizing Iterations.

Additional Attack on Surrogate Forensic Detector. Additionally, we include the table from Tab. 1 in the main paper, where we

Table 3: Performance of our method using different Diffusion Steps and Optimizing Iterations. The vertical E [17], R [8], D [18], S [9] are used to generate adversarial samples, The horizontal E [17], R [8], D [18], S [9] are used to detect adversarial samples.

	Diffusion Steps	Optimizing iterations	E	R	D	S
E	2	1	99.38	73.63	49.75	64.13
	2	5	100.00	89.50	66.38	84.13
	2	10	100.00	89.50	66.38	84.00
	3	1	99.38	70.63	48.50	63.88
	3	5	100.00	89.50	66.13	84.13
	3	10	100.00	89.50	66.50	84.00
	4	1	99.38	70.50	42.63	64.00
	4	5	100.00	88.25	65.88	83.25
	4	10	100.00	89.13	66.00	83.50
	2	1	59.25	99.75	59.38	86.00
	2	5	94.75	100.00	67.25	95.25
	2	10	94.75	100.00	67.13	95.25
R	3	1	59.38	99.75	59.50	84.75
	3	5	94.00	100.00	64.00	95.25
	3	10	94.13	100.00	63.88	95.25
	4	1	59.25	99.75	59.50	84.00
	4	5	94.13	100.00	64.13	95.25
	4	10	94.00	100.00	64.00	95.25
	2	1	81.25	64.13	99.38	94.25
	2	5	90.75	89.88	100.00	97.88
	2	10	90.75	89.88	100.00	97.88
	3	1	81.13	63.38	99.38	92.13
	3	5	89.88	88.13	100.00	97.38
	3	10	90.13	88.88	100.00	97.50
D	4	1	80.50	65.00	99.13	93.38
	4	5	89.88	87.13	100.00	96.75
	4	10	91.75	88.13	100.00	96.75
	2	1	46.50	54.88	43.13	99.68
	2	5	77.00	73.25	63.00	98.13
	2	10	77.00	73.25	62.50	98.25
	3	1	46.50	54.88	42.50	98.50
	3	5	76.63	73.50	62.13	98.00
	3	10	76.88	73.25	62.50	98.50
	4	1	46.50	54.88	42.25	98.38
	4	5	76.50	72.25	62.50	99.00
	4	10	77.00	72.88	62.88	99.50

use Efficientnet-B0 (E) [17], ResNet50 (R) [8], DeiT (D) [18], and Swin-T (S) [9] as detection models to detect adversarial examples generated using Efficientnet-B0 (E) [17], ResNet50 (R) [8], DeiT (D) [18], and Swin-T (S) [9] as surrogate models. As shown in Tab. 4 and Tab. 5, our method outperforms the baseline attack methods in both white-box and black-box settings, using various generation methods and different backbones as surrogate models.

3 ADDITIONAL QUALITY COMPARISONS

As shown in Fig. 1, we provide more quality comparisons of the adversarial samples generated with the baseline attacks *i.e.* FGSM [6], PGD [11], AutoAttack(AA) [4], DiffAttack [3] and Diff-PGD [22]. It is evident that traditional gradient-based transfer attack

methods [4, 6, 11] introduce visible noise patterns, whereas our diffusion model-based attack method produces images without noticeable noise patterns.

Table 4: The performance of attack methods evaluated using the Attack Success Rate, with the first column representing the methods EfficientNet-B0(E) [17], ResNet-50(R) [8] used to generate adversarial samples. The second column representing the methods EfficientNet-B0(E) [17], ResNet-50(R) [8], DeiT(D) [18], Swin-T(S) [9] used to detect adversarial samples. The third column represents different baseline attack methods FGSM [6], PGD [11], AutoAttack(AA) [4], Diff-PGD [22], DiffAttack [3], and our method. The first row represents different datasets, covering 8 sub-datasets in the GenImage dataset [23]: ADM [5], BigGAN [1], Glide [13], Midjourney [12], Stable Diffusion 1.4&1.5 [15], VQDM [7], Wukong [21]. Higher metric values indicate better performance, with the best results highlighted in bold.

		ADM	BigGAN	Glide	Midjourney	SDv14	SDv15	VQDM	Wukong	Average
E	E	FGSM	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		PGD	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		AA	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		DiffAttack	88.00	96.00	91.00	91.00	96.00	96.00	94.00	92.13
		DiffPGD	100.00	100.00	99.00	99.00	100.00	100.00	99.00	99.63
		Ours	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	R	FGSM	26.00	44.00	80.00	91.00	71.00	73.00	41.00	61.00
		PGD	38.00	68.00	90.00	95.00	93.00	94.00	42.00	74.50
		AA	37.00	43.00	83.00	94.00	94.00	92.00	41.00	69.38
		DiffAttack	41.00	60.00	54.00	72.00	70.00	80.00	87.00	67.50
		DiffPGD	42.00	76.00	30.00	56.00	77.00	76.00	90.00	66.00
		Ours	53.00	80.00	95.00	99.00	99.00	100.00	96.00	89.50
	D	FGSM	18.00	13.00	37.00	81.00	27.00	30.00	10.00	29.50
		PGD	36.00	32.00	59.00	89.00	45.00	43.00	30.00	46.75
		AA	40.00	25.00	58.00	92.00	53.00	44.00	35.00	47.63
		DiffAttack	49.00	31.00	40.00	76.00	39.00	44.00	62.00	46.63
		DiffPGD	55.00	55.00	29.00	43.00	32.00	33.00	73.00	43.88
		Ours	60.00	50.00	83.00	98.00	68.00	69.00	52.00	66.38
	S	FGSM	68.00	93.00	98.00	99.00	93.00	89.00	73.00	86.38
		PGD	99.00	100.00	100.00	100.00	100.00	100.00	95.00	98.75
		AA	99.00	74.00	100.00	98.00	99.00	99.00	93.00	93.88
		DiffAttack	43.00	58.00	65.00	62.00	57.00	53.00	68.00	57.00
		DiffPGD	30.00	69.00	19.00	13.00	17.00	17.00	32.00	27.25
		Ours	74.00	80.00	91.00	93.00	88.00	84.00	77.00	84.13
R	E	FGSM	92.00	71.00	99.00	98.00	94.00	87.00	57.00	83.88
		PGD	99.00	74.00	100.00	98.00	99.00	98.00	81.00	88.00
		AA	97.00	70.00	98.00	95.00	99.00	96.00	38.00	84.63
		DiffAttack	43.00	64.00	48.00	62.00	43.00	39.00	62.00	49.00
		DiffPGD	48.00	62.00	16.00	63.00	50.00	55.00	35.00	60.00
		Ours	99.00	99.00	100.00	100.00	99.00	98.00	68.00	94.75
	R	FGSM	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		PGD	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		AA	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		DiffAttack	81.00	81.00	70.00	94.00	97.00	97.00	78.00	85.75
		DiffPGD	100.00	100.00	100.00	98.00	100.00	100.00	100.00	99.75
		Ours	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	D	FGSM	19.00	16.00	35.00	70.00	22.00	29.00	3.00	25.88
		PGD	58.00	53.00	78.00	89.00	51.00	56.00	40.00	57.75
		AA	62.00	42.00	73.00	87.00	46.00	48.00	17.00	51.00
		DiffAttack	47.00	39.00	34.00	39.00	13.00	18.00	52.00	32.38
		DiffPGD	57.00	45.00	20.00	35.00	33.00	29.00	70.00	39.13
		Ours	68.00	61.00	85.00	97.00	68.00	67.00	42.00	67.25
	S	FGSM	85.00	98.00	98.00	100.00	94.00	91.00	78.00	82.00
		PGD	100.00	100.00	100.00	100.00	100.00	99.00	98.00	99.38
		AA	100.00	94.00	99.00	98.00	100.00	98.00	45.00	90.75
		DiffAttack	52.00	59.00	49.00	39.00	31.00	32.00	68.00	46.63
		DiffPGD	27.00	62.00	18.00	14.00	13.00	21.00	32.00	25.38
		Ours	93.00	97.00	100.00	100.00	99.00	98.00	86.00	95.25

Table 5: The performance of attack methods evaluated using the Attack Success Rate, with the first column representing the methods DeiT(D) [18], Swin-T(S) [9] used to generate adversarial samples. The second column representing the methods EfficientNet-B0(E) [17], ResNet-50(R) [8], DeiT(D) [18], Swin-T(S) [9] used to detect adversarial samples. The third column represents different baseline attack methods FGSM [6], PGD [11], AutoAttack(AA) [4], Diff-PGD [22], DiffAttack [3], and our method. The first row represents different datasets, covering 8 sub-datasets in the GenImage dataset [23]: ADM [5], BigGAN [1], Glide [13], Midjourney [12], Stable Diffusion 1.4&1.5 [15], VQDM [7], Wukong [21]. Higher metric values indicate better performance, with the best results highlighted in bold.

			ADM	BigGAN	Glide	Midjourney	SDv14	SDv15	VQDM	Wukong	Average
D	E	FGSM	68.00	61.00	90.00	98.00	85.00	81.00	34.00	69.00	73.25
		PGD	80.00	66.00	97.00	98.00	89.00	91.00	68.00	79.00	83.50
		AA	78.00	68.00	96.00	95.00	92.00	91.00	59.00	78.00	82.13
		DiffAttack	59.00	64.00	61.00	91.00	64.00	75.00	66.00	59.00	67.38
		DiffPGD	46.00	67.00	16.00	56.00	55.00	60.00	31.00	57.00	48.50
		Ours	86.00	97.00	96.00	100.00	96.00	94.00	66.00	91.00	90.75
	R	FGSM	37.00	46.00	79.00	88.00	76.00	68.00	40.00	61.00	61.88
		PGD	41.00	53.00	86.00	95.00	89.00	88.00	41.00	67.00	70.00
		AA	46.00	53.00	83.00	92.00	88.00	82.00	42.00	70.00	69.50
		DiffAttack	75.00	78.00	76.00	81.00	79.00	78.00	95.00	69.00	78.88
		DiffPGD	42.00	73.00	20.00	48.00	73.00	71.00	85.00	75.00	60.88
		Ours	69.00	71.00	95.00	100.00	97.00	99.00	96.00	92.00	89.88
	D	FGSM	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		PGD	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		AA	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		DiffAttack	98.00	96.00	91.00	91.00	79.00	81.00	93.00	63.00	86.50
		DiffPGD	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		Ours	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	S	FGSM	93.00	98.00	98.00	100.00	97.00	97.00	81.00	89.00	94.13
		PGD	100.00	100.00	100.00	100.00	100.00	100.00	95.00	97.00	99.00
		AA	100.00	100.00	98.00	97.00	100.00	99.00	95.00	97.00	98.25
		DiffAttack	71.00	62.00	66.00	80.00	80.00	79.00	87.00	56.00	72.63
		DiffPGD	24.00	57.00	20.00	16.00	14.00	20.00	33.00	23.00	25.88
		Ours	99.00	95.00	100.00	100.00	100.00	99.00	95.00	95.00	97.88
S	E	FGSM	32.00	38.00	62.00	89.00	59.00	64.00	22.00	46.00	51.50
		PGD	45.00	50.00	53.00	84.00	51.00	52.00	30.00	43.00	51.00
		AA	38.00	39.00	43.00	77.00	54.00	53.00	27.00	39.00	46.25
		DiffAttack	7.00	30.00	17.00	26.00	14.00	23.00	53.00	24.00	24.25
		DiffPGD	34.00	56.00	12.00	39.00	44.00	44.00	29.00	45.00	37.88
		Ours	59.00	82.00	81.00	97.00	87.00	86.00	45.00	79.00	77.00
	R	FGSM	12.00	5.00	33.00	80.00	57.00	50.00	33.00	43.00	39.13
		PGD	14.00	29.00	52.00	89.00	77.00	75.00	38.00	62.00	54.50
		AA	10.00	23.00	55.00	91.00	81.00	79.00	39.00	61.00	54.88
		DiffAttack	11.00	15.00	18.00	28.00	40.00	31.00	66.00	31.00	30.00
		DiffPGD	32.00	65.00	22.00	43.00	65.00	60.00	85.00	63.00	54.38
		Ours	22.00	41.00	54.00	95.00	94.00	97.00	93.00	90.00	73.25
	D	FGSM	18.00	11.00	35.00	83.00	32.00	36.00	12.00	26.00	31.63
		PGD	35.00	26.00	55.00	87.00	33.00	32.00	20.00	25.00	39.13
		AA	37.00	30.00	54.00	86.00	41.00	41.00	24.00	33.00	43.25
		DiffAttack	33.00	25.00	22.00	30.00	23.00	22.00	51.00	21.00	28.38
		DiffPGD	42.00	42.00	17.00	29.00	24.00	24.00	67.00	20.00	33.13
		Ours	56.00	50.00	72.00	97.00	66.00	61.00	51.00	51.00	63.00
	S	FGSM	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		PGD	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		AA	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		DiffAttack	79.00	96.00	91.00	64.00	76.00	82.00	80.00	76.00	80.50
		DiffPGD	56.00	89.00	57.00	52.00	71.00	63.00	75.00	62.00	65.63
		Ours	100.00	93.00	98.00	100.00	99.00	99.00	96.00	100.00	98.13

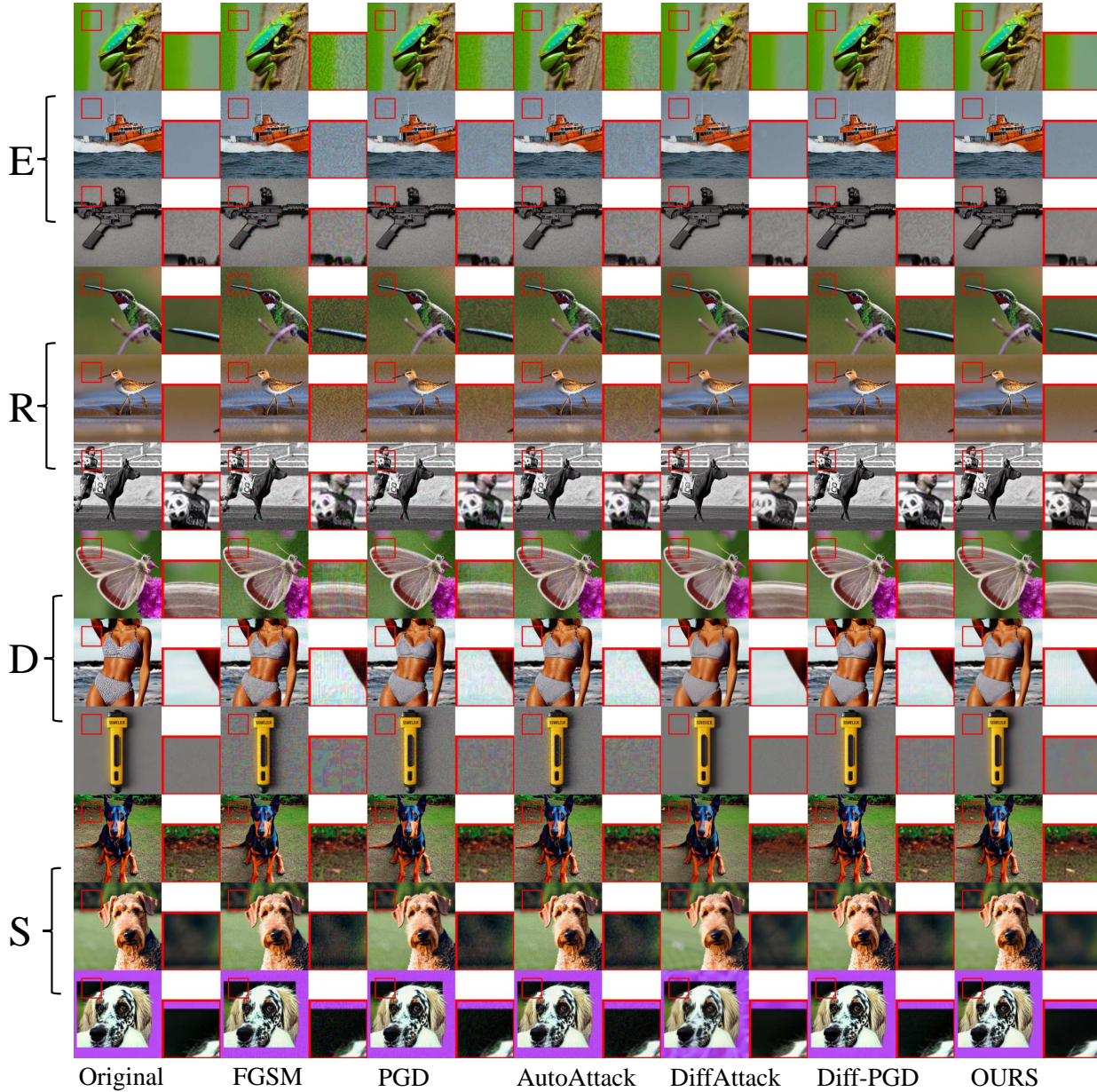


Figure 1: Additional Qualitative assessment of adversarial examples generated by FGSM [6], PGD [11], AutoAttack(AA) [4], DiffAttack [3], Diff-PGD [22], and our method on the GenImage dataset [23]. These samples were generated from different backbones, namely EfficientNet-B0(E) [17], ResNet-50(R) [8], DeiT(D) [18] and Swin-T(S) [9].

REFERENCES

- [1] Andrew Brock, Jeff Donahue, and Karen Simonyan. 2018. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096* (2018).
- [2] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. 2022. End-to-End Reconstruction-Classification Learning for Face Forgery Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4113–4122.
- [3] Jianqi Chen, Hao Chen, Keyan Chen, Yilan Zhang, Zhengxia Zou, and Zhenwei Shi. 2023. Diffusion Models for Imperceptible and Transferable Adversarial Attack. *arXiv preprint arXiv:2305.08192* (2023).
- [4] Francesco Croce and Matthias Hein. 2020. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*. PMLR, 2206–2216.
- [5] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* 34 (2021), 8780–8794.
- [6] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).
- [7] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. 2022. Vector Quantized Diffusion Model for Text-to-Image Synthesis. *arXiv:2111.14822 [cs.CV]*

- [8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [9] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*. 10012–10022.
- [10] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. 2021. Generalizing face forgery detection with high-frequency features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16317–16326.
- [11] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* (2017).
- [12] MidJourney. 2022. <https://www.midjourney.com/home/>.
- [13] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741* (2021).
- [14] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. 2023. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 24480–24489.
- [15] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.
- [16] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. 2023. Learning on Gradients: Generalized Artifacts Representation for GAN-Generated Images Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12105–12114.
- [17] Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*. PMLR, 6105–6114.
- [18] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*. PMLR, 10347–10357.
- [19] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. 2020. CNN-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8695–8704.
- [20] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. 2023. DIRE for Diffusion-Generated Image Detection. *arXiv preprint arXiv:2303.09295* (2023).
- [21] Wukong. 2022. <https://xihe.mindspore.cn/modelzoo/wukong>.
- [22] Haotian Xue, Alexandre Araujo, Bin Hu, and Yongxin Chen. 2023. Diffusion-Based Adversarial Sample Generation for Improved Stealthiness and Controllability. *arXiv preprint arXiv:2305.16494* (2023).
- [23] Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. 2023. GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image. *arXiv preprint arXiv:2306.08571* (2023).