Spatial-Temporal Forgery Trace based Forgery Image Identification

Supplementary Material

8. Appendix

This appendix provides supplementary details on the research methodology and experimental results. First, we present the performance of the proposed method in detecting forged images and provide related analysis. Next, we present the overall experimental results, comparing our method with 12 mainstream forgery detection models and STFT with mainstream generative image detection methods. Finally, we conduct an in-depth analysis of the experimental results, discussing the performance of different methods and highlighting the advantages of our model in various tasks. The source code is available publicly at https://github.com/GCLion/STFT.

8.1. Comparative Performance Analysis of STFT

Table 4 presents a comparative analysis of the accuracy (ACC, %) of the proposed STFT method against state-of-the-art (SOTA) forgery detection models across various image generators. As observed, STFT achieves the highest average accuracy of 94.14%, significantly outperforming existing methods. Notably, it achieves near-perfect detection on Wukong (99.99%), SDv1.4 (99.65%), and SDv1.5 (99.99%), demonstrating exceptional efficacy in detecting forgeries generated by diffusion-based models. Additionally, STFT also outperforms all competing methods on GLIDE (96.28%), Midjourney (95.63%), and VQDM (93.87%), further highlighting its ability to effectively capture forgery traces across various diffusion models, thereby exhibiting strong generalization capability. However, the detection performance is relatively lower on BigGAN (84.13%) and ADM (83.61%). Since BigGAN is a non-diffusion model (GAN-based), its forgery characteristics differ from those of diffusion-based models, making detection more challenging. Similarly, ADM-generated images may contain more complex forgery artifacts, increasing detection difficulty. This suggests that STFT could be further optimized, particularly in detecting forgeries produced by GAN-based models, to improve robustness across a broader range of forgery techniques. Overall, STFT leverages spatio-temporal forgery trace modeling and frequency-domain enhancement, achieving state-of-the-art performance on diffusion-based forgery detection while maintaining strong generalization across various datasets.

Table 5 compares the AUC (%) performance of the STFT method with existing state-of-the-art forgery detection methods on the DeepFaceGen dataset across different generative models. As shown in the table, STFT achieves the highest average AUC across all test sets, reaching 95.21%, significantly outperforming all comparative methods. STFT demonstrates superior performance across multiple datasets, including VD (99.03%), DF-GAN (99.24%), and OJ (98.87%), showcasing its outstanding detection capability. Furthermore, STFT outperforms other methods on datasets such as Midjourney (94.73%), DALL·E 1 (90.91%), DALL·E 3 (92.02%), Wenxin (93.65%), and SD1 (94.08%), further validating its strong generalization ability across different text-to-image generation models. However, on the SDXL (96.14%) and SD2 (95.91%) datasets, STFT is not the absolute best. The highest score on SDXL is achieved by RECCE (96.75%), while SD2's highest score is close to STFT's, with a slight margin. This discrepancy may be due to the more complex high-resolution features present in images generated by SDXL and SD2 during training, with certain traditional methods (such as RECCE) being more sensitive to these features. Overall, STFT establishes a leading advantage across various forgery generation models on DeepFaceGen through spatiotemporal forgery trace modeling and frequency domain enhancement, achieving optimal performance on the majority of datasets. This confirms its excellent generalization and adaptability.

8.2. Failure Case Visualization and Analysis

Figure 4 presents a collection of misclassified images. We analyze them from the temporal, spatial, and frequency perspectives of the model.

The STFT method relies on the temporal distribution features of generative diffusion models to detect forged images. However, in some misclassified samples, the temporal distribution patterns of certain forged images closely resemble those of real images, making them difficult to distinguish. For example, high-quality images generated by DALL·E 3 and Midjourney often closely align with real data distributions, interfering with the model's ability to differentiate them.

STFT assumes that forged images exhibit abnormal correlations in temporal features across different spatial regions—being either highly similar or completely different. However, in misclassified samples, some generated images display natural feature transitions, making spatial forgery traces harder to detect. For instance, images generated by Midjourney maintain a high level of consistency in facial details and textures, reducing the effectiveness of STFT in identifying spatial forgery artifacts.

STFT employs high-frequency components as weighting factors to enhance forgery detection. However, in misclassified samples, we observed that some forged images ex-

Method	Year	Midjourney	SDv1.4	SDv1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	Avg.
CNNSpot [34]	2021	84.92	99.88	99.76	53.48	53.80	99.68	55.50	49.93	74.62
F3Net [33]	2020	77.85	98.99	99.08	51.20	54.87	97.92	58.99	49.21	73.51
CLIP/RN50 [34]	2021	83.30	99.97	99.89	54.55	57.37	99.52	57.90	50.00	75.31
GramNet [24]	2020	73.68	98.85	98.79	51.52	55.38	95.38	55.15	49.41	72.27
De-fake [39]	2023	79.88	99.86	99.62	68.62	71.57	98.42	78.43	74.37	84.73
Conv-B [26]	2022	83.55	99.99	99.92	51.75	56.27	99.91	58.41	50.00	74.98
Swin-T [25]	2021	62.11	**99.99**	99.88	49.85	67.62	99.01	62.28	57.63	74.79
UnivFD [31]	2023	91.46	96.41	96.14	58.07	73.40	94.53	67.83	57.72	79.45
DIRE [45]	2023	50.40	99.99	99.92	52.32	67.23	99.98	50.10	49.99	71.24
PatchCraft [52]	2023	79.00	89.50	89.30	77.30	78.40	89.30	83.70	72.40	82.30
AIDE [49]	2024	79.38	99.74	99.76	78.54	91.82	98.65	80.26	66.89	86.88
DRCT [6]	2024	91.50	95.01	94.41	79.42	89.18	94.67	90.03	81.67	89.49
STFT (Ours)	-	**95.63**	99.65	**99.99**	**83.61**	**96.28**	**99.99**	**93.87**	**84.13**	**94.14**

Table 4. Accuracy (ACC, %) comparison of our STFT method and other forgery detection models across various image generators. All methods were trained on GenImage/SDv1.4 and evaluated on different test subsets.

Generator	Year	Xception	EfficientNet	F3Net	RECCE	DNADet	FreqNet	DIRE	DRCT	UnivFD	NPR	STFT (Ours)
Midjourney [21]	2022	77.01	79.52	81.65	87.64	93.44	85.69	87.01	89.78	88.67	89.01	**94.73**
DALL·E 1 [35]	2021	75.45	81.74	84.73	83.25	85.62	84.25	86.65	89.91	87.64	89.54	**90.91**
DALL·E 3 [32]	2023	86.59	88.41	87.23	89.17	83.90	87.13	87.84	88.05	89.21	89.41	**92.02**
Wenxin [1]	2023	86.87	84.34	89.28	92.13	91.60	91.98	92.84	92.72	90.01	92.35	**93.65**
SD1 [38]	2022	87.64	86.83	90.95	89.83	92.40	90.94	92.35	93.56	90.01	90.12	**94.08**
SDXLR [40]	2023	84.13	93.46	86.40	90.46	89.03	89.40	88.61	89.51	89.01	88.64	**94.32**
OJ [15]	2024	89.72	89.00	92.72	96.90	94.04	93.08	91.28	92.45	88.01	90.28	**98.87**
pix2pix [20]	2017	83.42	77.61	81.21	89.71	88.52	88.66	89.01	91.51	89.54	89.30	**93.63**
SD2 [40]	2023	87.79	85.91	89.11	95.43	92.70	90.92	89.54	90.41	91.45	90.01	**95.91**
SDXL [38]	2023	86.06	87.65	91.43	**96.75**	92.28	92.61	91.54	91.01	90.01	89.87	96.14
VD [30]	2023	85.02	83.84	89.52	95.67	89.21	96.55	98.78	94.25	89.68	91.01	**99.03**
DF-GAN [22]	2021	95.42	96.71	93.45	93.54	94.22	97.11	90.32	92.54	95.01	98.88	**99.24**
Average	-	85.42	86.25	88.14	91.70	90.58	90.69	90.69	90.48	91.30	89.85	**95.21**

Table 5. Performance Comparison (AUC, %) of STFT and other methods across various forgery generators on DeepFaceGen.

hibit prominent high-frequency components due to complex textures, leading to false positives. Additionally, some real images undergo post-processing techniques such as denoising and sharpening, which weaken high-frequency features, making detection more challenging. For example, images generated by ADM contain fewer high-frequency forgery traces, impacting the model's detection performance.

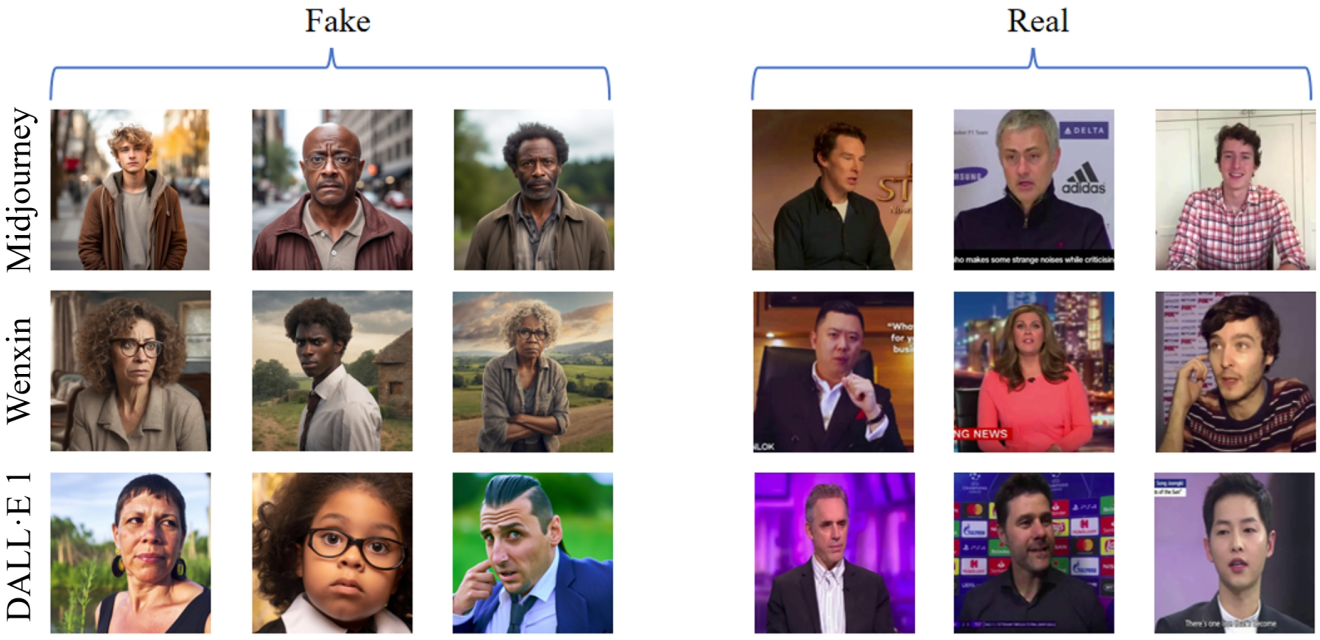

Figure 4. Some failure cases. The left side presents high-realism forged images synthesized by Midjourney, Wenxin, and DALL·E 1-based generative models, which can be easily mistaken for real photographs. The right side shows real images for comparison.