A Appendix

A.1 Experimental Details

Dataset. The experiments are conducted using the CelebA-HQ dataset [6]. 300 standard samples that meet the requirements for successful face localization are selected to verify the performance of DFPD. Since the proposed DFPD method also needs to validate the performance of proactive defense against traditional manipulations, it is necessary to perform tampering on the obtained defensed samples. Based on this, this paper uses Photoshop to create a standard test environment against traditional tampering. Given the high incidence of splicing and copy-move attacks in face forgery, the tampering focuses on simulating these two types of tampering to align the test scenarios with real attack scenarios. In order to prevent unfair comparisons caused by the perturbation of the DFPD method to the compared methods, homologous tampering is applied to both original and defensed images. Then, the tampered original images are used for the compared traditional passive localization methods, while the tampered defensed images are for the proposed method.

Implementation Details. To comprehensively evaluate the DFPD's performance in deepfakes defense, this paper selects four representative deepfake models: SimSwap [1], FaceShifter [8], FGAN [11], and StarGAN [3]. These models cover the two main forgery techniques: face swapping and face attribute editing. SimSwap and FaceShifter focus on high-fidelity face swapping, while FGAN and StarGAN handle face attribute editing. Testing on these models verifies the framework's effectiveness and universality against different forgery types. For each test image, five attributes are selected: "black hair," "blond hair," "brown hair," "gender," and "age." For face swapping, a target face identity is randomly selected and swapped onto the test image, i.e., the test image is used as the source face for defense operations.

In the perturbation adding module, the perturbation intensity threshold is set to 0.02, the iteration step size α is 0.002, the batch size is 1, the number of iterations is 20, and the weights of the forgery models are set to 1 except for w_i corresponding to the SimSwap model, which is set to 2. In the MPP module, the size of K is 3×3, the threshold of T is 300, and τ is set to 0.2.

Baselines. Previous methods failed to achieve the dual-forgery defense function against deepfakes and traditional manipulations. Therefore, the proposed DFPD method is compared with existing methods in terms of proactive defense against deepfakes and passive tampering localization against traditional manipulations, respectively.

For proactive defense performance evaluation against deepfakes, the DD method, FOUND method, and SA method are selected. The DD method [13] embeds robust watermarks to disrupt face-swapping and enable traceability. The FOUND method [12] uses a two-stage strategy and gradient ensemble to generate universal defensed samples. The SA method [9] generates a defensed image by adding perturbations to the face region. Additionally, modified versions of these methods, the DD (Ensemble) and SA (Ensemble), are also included, which replace the single proxy model with a pool of proxy models composed of four forged models considered by the proposed method. For a fair comparison, some methods are retrained separately. The DD and DD (Ensemble) methods omit the watermark extraction module and its loss function and use the difficulty model SimSwap model for retraining. The FOUND method replaces the ensemble models with the four forgery models of this paper while retaining the original settings.

For proactive tampering localization performance evaluation against traditional manipulations, four mainstream detection methods are selected: MVSS-Net [2] combines multi-view features and multi-scale supervision to learn tamper-sensitive features while reducing false alarms; PSCC-Net [10] uses top-down feature extraction and bottom-up mask estimation with a space-channel correlation module for robust detection; EITLNet [5] enhances feature representation and fuses RGB and noise features at multiple scales to improve localization accuracy; TruFor [4] extracts RGB information and learns noise-sensitive fingerprints using a Transformer-based architecture for reliable detection.

Evaluation Metrics. For visual quality assessment of defensed samples, PSNR, SSIM, and LPIPS are used. For the defense performance evaluation of deepfakes, PSNR, LPIPS, and L_1 norm are applied. For the tampering localization performance evaluation of traditional manipulations, IoU, F1 score, AUC, and ACC are utilized.

A.2 Cross-Dataset Generalizability

To further evaluate the effectiveness of the proposed DFPD method, its cross-dataset generalizability is tested. Specifically, to assess cross-racial generalizability, 100 Asian faces are selected from FFHQ [7] for testing. The performance of DFPD on this FFHQ-Asian subset and the original CelebA-HQ dataset is compared in Table 1. The metrics for both deepfake defense and traditional tampering localization remain strong across both datasets, which indicates that the DFPD method exhibits good cross-dataset generalizability.

A.3 Robustness Analysis

To assess the durability of the protection offered by DFPD in practical scenarios, its robustness against common image post-processing operations is evaluated. Such operations may be performed by an attacker to remove the defensive protection, or may occur through standard procedures like social media sharing. The evaluation considers three simulated attacks: a Gaussian blur with a filter kernel size of 5 and a standard deviation of 0.5; social media communication simulated by uploading images to Twitter and then downloading them for testing; and a resizing attack, which scales images by a factor of 1.5 using bicubic interpolation.

The performance of DFPD under these attacks is detailed in Table 2. While a degree of performance degradation is observed, the method's dual-defense capabilities remain largely intact. For deepfake defense, the disruptive effect on forged outputs is maintained, showing only a modest decline. Similarly, for traditional manipulation localization, the key metrics like F1 and AUC remain at a high level, indicating that tampered regions can still be effectively identified. These results suggest that DFPD possesses a certain level of robustness against common image post-processing.

Table 1: Performance comparison on CelebA-HQ and FFHQ-Asian datasets.

Datasets	Defense against Deepfakes			Defense against Traditional Manipulation			
	PSNR↓	LPIPS↑	L_1↑	AUC↑	F1↑	IoU↑	ACC↑
FFHQ-Asian	13.33	0.4005	0.43	0.9593	0.9113	0.8418	0.9777
CelebA-HQ	13.71	0.3835	0.41	0.9672	0.9391	0.8887	0.9837

Table 2: Robustness evaluation of DFPD under various attacks.

Attacks	Defense against Deepfakes			Defense against Traditional Manipulation			
	PSNR↓	LPIPS↑	L_1↑	AUC↑	F1↑	IoU↑	ACC↑
No attack	**13.71**	**0.3835**	**0.41**	**0.9672**	**0.9391**	**0.8887**	**0.9837**
GussianBlur	16.20	0.2967	0.31	0.9507	0.7832	0.6626	0.9317
SocialMedia	13.79	0.3807	0.40	0.9599	0.9259	0.8668	0.8789
Resizing	18.13	0.2544	0.22	0.9671	0.8768	0.7919	0.9663

Table 3: Comparison of computational time for generating a defensed sample.

Algorithms	Need Train	Time(s)
FOUND	✓	0.32
SA	×	3.50
SA(Ensemble)	×	6.70
DD	✓	0.23
DD(Ensemble)	✓	0.50
Ours	×	6.00

A.4 Computational Cost

In order to assess the feasibility of DFPD in practical applications, its computational cost is analyzed in this section. DFPD employs an iterative adversarial attack methodology, which, similar to SA and SA (Ensemble), does not require pre-training and can be used directly for inference. This is in contrast to methods such as FOUND and DD, which require pre-training of a generic perturbation or perturbation generator.

To quantify the runtime efficiency, an experiment is conducted to evaluate the time required to generate a single defensed sample. As shown in Table 3, the running time of DFPD is comparable to SA(Ensemble). The experimental results show that although DFPD integrates an additional watermark embedding step, its added complexity to achieve dual defense does not significantly increase the computational burden, demonstrating its potential for real-world deployment.

References

[1] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. 2020. SimSwap: An Efficient Framework For High Fidelity Face Swapping. In *Proceedings of the 28th ACM International Conference on Multimedia* (Seattle, WA, USA) *(MM '20)*. Association for Computing Machinery, New York, NY, USA, 2003–2011. doi:10.1145/3394171.3413630

[2] Xinru Chen, Chengbo Dong, Jiaqi Ji, Juan Cao, and Xirong Li. 2021. Image Manipulation Detection by Multi-View Multi-Scale Supervision . In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE Computer Society, Los Alamitos, CA, USA, 14165–14173. doi:10.1109/ICCV48922.2021.01392

[3] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 8789–8797. doi:10.1109/cvpr.2018.00916

[4] Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. 2023. TruFor: Leveraging All-Round Clues for Trustworthy Image Forgery Detection and Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20606–20615. https://doi.org/10.1109/CVPR52729.2023.01974

[5] Kun Guo, Haochen Zhu, and Gang Cao. 2024. Effective Image Tampering Localization Via Enhanced Transformer and Co-Attention Fusion. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 4895–4899. doi:10.1109/ICASSP48485.2024.10446332

[6] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive Growing of Gans for Improved Quality, Stability, and Variation. *arXiv preprint arXiv:.10196* (2018). doi:10.48550/arXiv.1710.10196

[7] Tero Karras, Samuli Laine, and Timo Aila. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4396–4405. doi:10.1109/CVPR.2019.00453

[8] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. 2019. FaceShifter: Towards High Fidelity And Occlusion Aware Face Swapping. arXiv:1912.13457 http://arxiv.org/abs/1912.13457

[9] Qilei Li, Mingliang Gao, Guisheng Zhang, and Wenzhe Zhai. 2023. Defending Deepfakes by Saliency-Aware Attack. *IEEE Transactions on Computational Social Systems* (2023), 1–8. doi:10.1109/TCSS.2023.3271121

[10] Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. 2022. PSCC-Net: Progressive Spatio-Channel Correlation Network for Image Manipulation Detection and Localization. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 11 (Nov. 2022), 7505–7517. doi:10.1109/TCSVT.2022.3189545

[11] Md Mahfuzur Rahman Siddiquee, Zongwei Zhou, Nima Tajbakhsh, Ruibin Feng, Michael B. Gotway, Yoshua Bengio, and Jianming Liang. 2019. Learning Fixed Points in Generative Adversarial Networks: From Image-to-Image Translation to Disease Detection and Localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 191–200. doi:10.1109/iccv.2019.00028

[12] Long Tang, Dengpan Ye, Zhenhao Lu, Yunming Zhang, and Chuanxi Chen. 2024. Feature Extraction Matters More: An Effective and Efficient Universal Deepfake Disruptor. *ACM Trans. Multimedia Comput. Commun. Appl.* 21, 2, Article 46 (Dec. 2024), 22 pages. doi:10.1145/3653457

[13] Yunming Zhang, Dengpan Ye, Caiyun Xie, Long Tang, Xin Liao, Ziyi Liu, Chuanxi Chen, and Jiacheng Deng. 2024. Dual Defense: Adversarial, Traceable, and Invisible Robust Watermarking Against Face Swapping. *IEEE Transactions on Information Forensics and Security* 19 (2024), 4628–4641. doi:10.1109/TIFS.2024.3383648