Watermarking Matters for Deepfake Detection: A Proactive Method for Detecting Forgeries under Conventional Attacks

Supplementary Material

Due to space limitations in the main text, additional experimental descriptions and results are provided in the supplementary materials. The supplementary results cover two aspects: parameter analysis and visualization.

For each facial image to be protected, five randomly generated 16×16 binary watermarks are embedded during both the training and testing phases. Accordingly, all experimental results are reported as the average over five test runs. The source code will be made publicly available.

1. Parameter Analysis

1.1. Analysis of the Selection of n_{max}

In Eq. (3), $n_{\max}$ and $m_{\max}$ represent the highest order and repetition of QCOMs, respectively. These two parameters influence the imperceptibility, robustness, and computational efficiency of the watermarking process. In our implementation, we set $n_{\max}$=$m_{\max}$ and explore how varying $n_{\max}$ affects performance. The experimental results on the CelebA dataset are shown in Fig. 1. As $n_{\max}$ increases, both PSNR and computation time rise, while BER initially decreases and then stabilizes. To balance computational efficiency and watermarking performance, we set $n_{\max}$=50.

1.2. Analysis of the Selection of Δ

In Eq. (4), we design a watermark embedding formula, where Δ denotes the embedding strength, which controls the trade-off between robustness and imperceptibility of watermark. In Fig. 2, we investigate the effect of different Δ values on the CelebA dataset. It can be observed that as Δ increases, the BER against both known and unknown conventional attacks decreases, indicating improved robustness. However, the PSNR decreases, suggesting that the watermark becomes more perceptible. To balance robustness and imperceptibility of watermark, we set Δ=0.5.

1.3. Analysis of the Number of SFmamba Blocks

The feature extraction backbone is constructed by cascading multiple SFmamba blocks. As presented in Table 1, we investigate how the number of SFmamba blocks affects watermark extraction. The experiments reveal that when the number of blocks exceeds six, the BER remains stable against both conventional and deepfake attacks. As shown in Fig. 3, further analysis indicates that the increase in ACC slows down beyond six blocks. To balance performance and computational cost optimall, we finally adopt six SFmamba blocks to construct the backbone network.

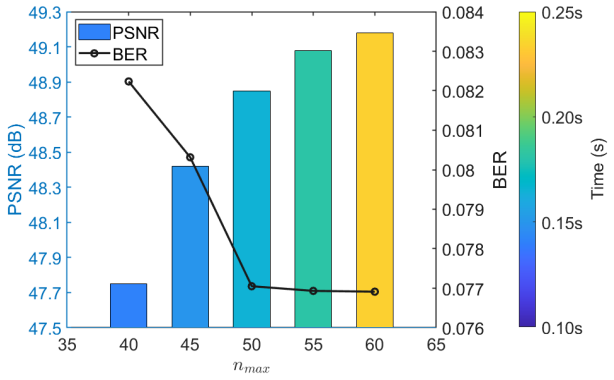

Figure 1. The impact of different n_{max} values on watermark embedding time, imperceptibility, and robustness against conventional attacks.

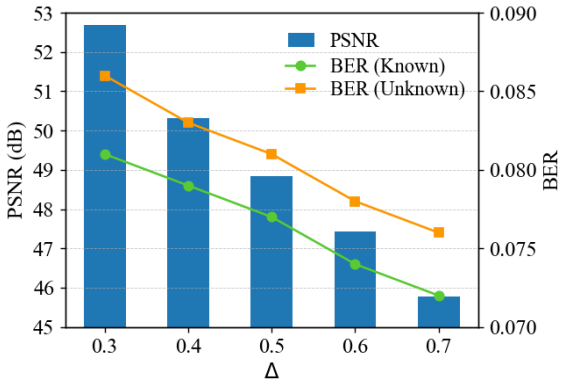

Figure 2. The impact of different Δ values on the robustness and imperceptibility of the watermark.

Table 1. The impact of the number of SFmamba blocks on watermark extraction.

Number of blocks	4	5	**6**	7	8
Conventional attack	8.250	8.246	8.243	8.242	8.241
Deepfake attack	89.416	89.420	89.423	89.424	89.422

2. Visualization Result

2.1. Visualization of Extracted Watermark

In Fig. 4, we present the visualization of the watermark extraction results for a watermark "w". It can be observed that the proposed method successfully extracts a watermark that closely resembles the original under conventional attacks.

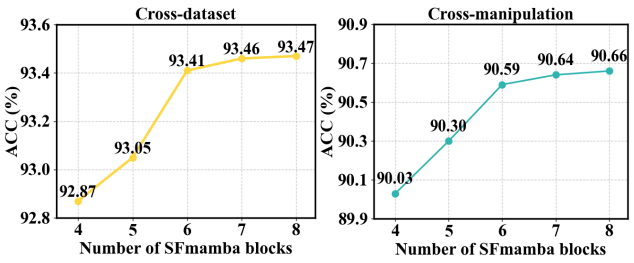

Figure 3. The impact of the number of SFmamba blocks on deepfake detection.

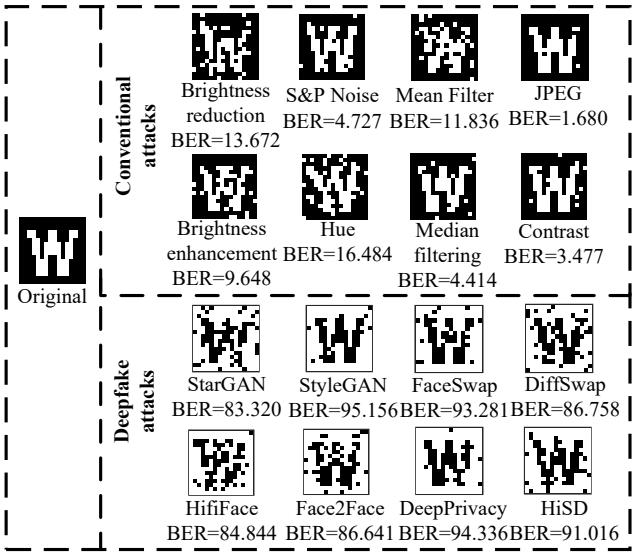

Figure 4. Visualization of watermark extraction under conventional and deepfake attacks.

In contrast, under deepfake attacks, the extracted watermark differs significantly from the original. These results demonstrate that the proposed method is robust against conventional attacks while exhibiting high sensitivity to deepfake manipulations, as the integrity of the watermark is substantially compromised.

2.2. Visualization of Feature Maps

Fig. 5 compares the attention regions learned by different models, including a deepfake detector trained independently and another optimized through our dual-task learning framework. During the testing phase, facial images subjected to deepfake attacks are fed into the models, and Class Activation Mapping (CAM) [1] is employed to visualize the feature maps. It can be seen that the dual-task learning framework focuses on critical facial features, particularly the eyes, nose, and mouth, leading its attention regions to align more closely with the actual manipulated areas than those of an independent deepfake detector. This improvement arises from the framework's effective utilization of reliable watermarking knowledge to uncover forgery clues. Although not all manipulated regions are completely covered, the attention distribution is sufficiently precise to enable the discriminator to accurately identify forged facial images.

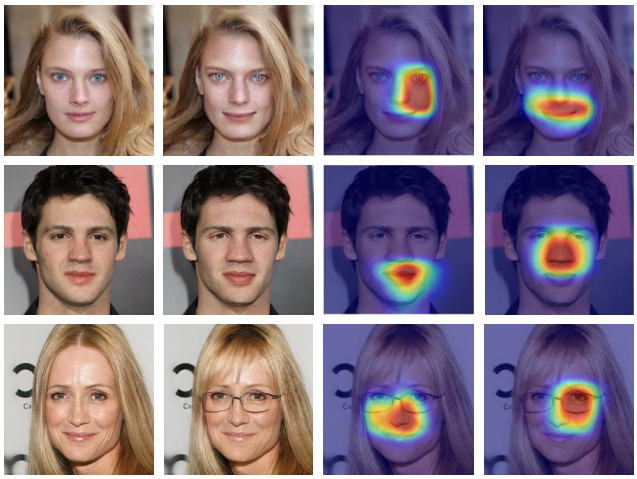

Figure 5. The visualization results of feature maps generated by CAM, from left to right, they are: real image, forged image, attention region obtained by training a deepfake detector independently, and attention region obtained by our dual-task learning framework.

References

[1] Lukang Wang, Min Zhang, and Wenzhong Shi. Cs-wscdnet: Class activation mapping and segment anything model-based framework for weakly supervised change detection. *IEEE Trans. Geosci. Remote. Sens.*, 61:1–12, 2023. 2