

Supplementary Materials: Advancing Generalized Deepfake Detector with Forgery Perception Guidance

Anonymous Authors

1 MODULE STRUCTURE

The extra explicit optimizations promote adapting sample diversities by considering the information on facial image qualities. Given a forgery sample, various image qualities can yield disparities within the forgery traces. It becomes important to adjust the detector to accommodate such quality variations. Therefore, we further incorporate an extra regression network branch and a lightweight attention module into the detector. Fig. 1 shows the details of the newly proposed structures. In the regression branch, there is only a group convolutional layer with 3×3 kernel size and a convolutional layer with 1×1 kernel size. Since the group number equals C , these two convolutional layers constitute the depthwise separable convolutional layer, in which the overall parameters and computational cost are low. Moreover, the parameters of the attention module are also lightweight. Similar to the Squeeze-and-Excitation module in [1], we focus on the channel dimension by averaging the spatial features in each channel. After that, the averaged channel features are respectively squeezed and enlarged through fully connected layers. The channel dimension of the squeezed feature equals 256. After the re-enlarged features go through the sigmoid function, the weights will multiply the backbone features and get the feature F .

2 APPLICABILITY ANALYSIS

Considering the prevalence of distorted images across social media, an ideal applicable deepfake detector should demonstrate resilience against these image distortions. To further analyze the applicability of our proposed method under different image distortions, we test the detector by inputting the images with different distortions. Following the work in [2], Fig 2(a) shows the applicability of the detector to different kinds of image distortion. With the increasing distortion levels, more information within the original images is either interfered with or lost. According to the detection results, it can be noticed that the deepfake detector with FPG is robust to saturation, contrast, and block-wise distortion, which demonstrates the applicability of the detector against color-level distortions. Also, the detector is robust to the early-level Gaussian blur and JPEG compression. The samples afflicted by severe blur and compression distortions are presented in Fig 2(b). Since the presence of distinctive textures within the forgery traces is essential for perception, the severe blur and compression not only result in the loss of color information but also interfere with excessive textures, which may be the reason why the detection results appear declined.

3 SALIENCY VISUALIZATION

As mentioned in the main text, unlike the baseline method which lacks the connection between forgery sample generation and forgery perception cultivation, FPG investigates the deficiencies of forgery perceptions and adopts a refinement strategy to pertinently train

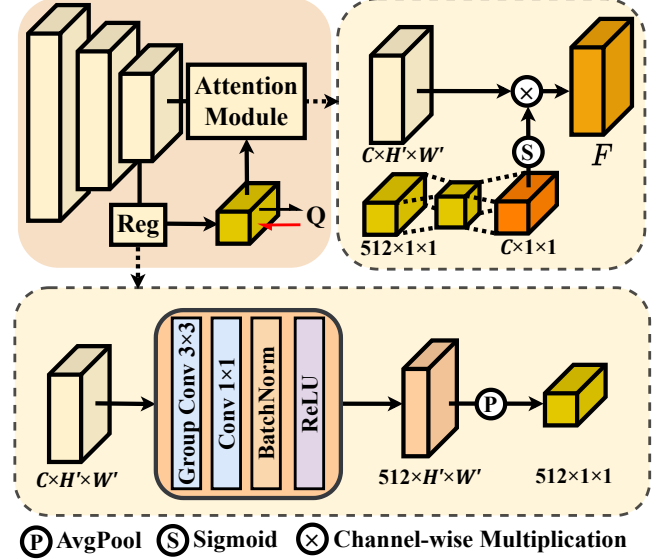


Figure 1: The details of the proposed regression network branch and the attention module. The output of the regression branch is then mapped to a value that is supervised by the output of the quality assessment network (denoted as Q).

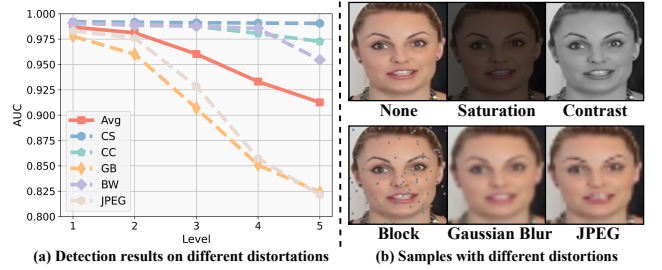


Figure 2: (a) Applicability analysis on the FF++ dataset under different image distortions. Here 'CS', 'CC', 'BW', 'GB', and 'JPEG' denote color saturation, color contrast, block-wise distortion, Gaussian blur, and JPEG compression, respectively. (b) Samples of the distortions considered at severity level 5.

the detector, thereby elevating the generalization efficiently. Moreover, FPG introduces more sample information as explicit optimizations, which makes the detector further adapt the sample diversities. To further investigate the perception of the detector to the forgery samples, more visualization examples from different datasets are shown in Fig 3. With the help of FPG, even if the detector unknowns these datasets, there are still higher salient values than the baseline

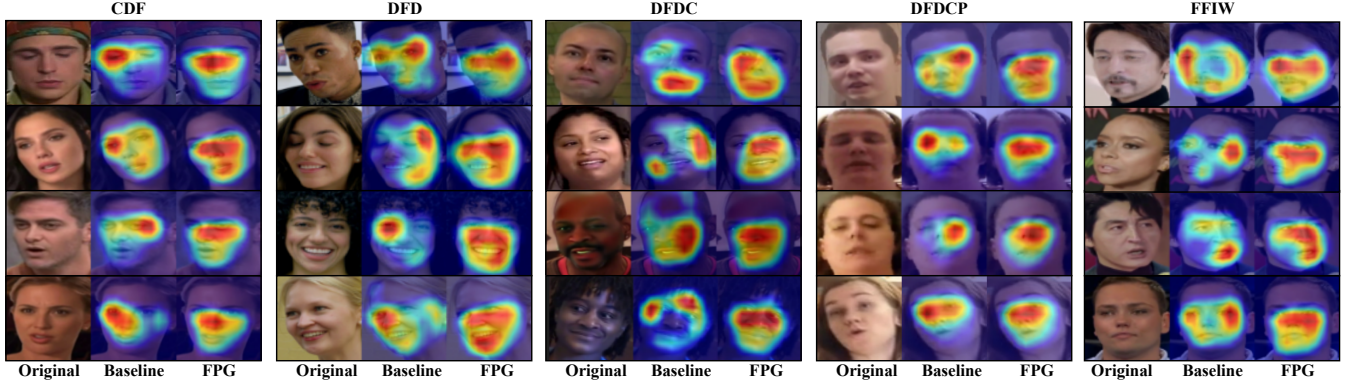


Figure 3: The saliency visualization of forgery samples from different unknown public datasets.

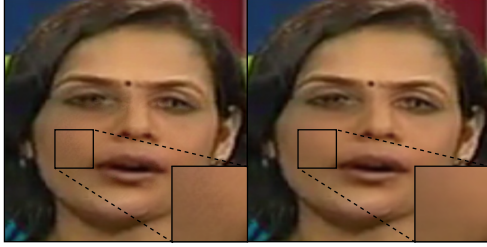


Figure 4: The refined forgery samples with different perturbations. The left is the sample with discrete perturbation. The right is the sample with unified perturbation.

Table 1: Ablation study of refinement sequence. ‘S’ and ‘M’ denote the forgery shape refinement and forgery magnitude refinement, respectively.

Sequence	Test Set AUC (%)					
	CDF	DFD	DFDC	DFDCP	FFIW	Avg
M → S	93.45	97.14	74.50	87.26	88.17	88.10
S → M	94.49	96.41	74.75	87.24	87.93	88.16

method within the forgery faces, which means the detector can perceive the forgery traces more completely.

4 MORE ABLATION STUDIES

Sample Refinement Sequence. In our refinement strategy, the sequence of sample refinement is first to refine the shape and then the magnitude of the forgery traces. To investigate the impact of the refinement sequence, we adjust the sequence by first refining the magnitude and then the shape of the forgery traces. The results under different datasets are shown in Table 1, it can be noticed that there is only a 0.06 % difference (88.16% vs. 88.10%) in the average detection results for different refinement sequences. This observation underscores the robustness of the refinement sequence.

Different Perturbation Settings. In forgery magnitude refinement, the forgery masks are added with the perturbations to enlarge

Table 2: Ablation study of different perturbation settings in forgery magnitude refinement.

Perturbation	Test Set AUC (%)			
	CDF	DFDCP	FFIW	Avg
Discrete	94.32	85.60	87.49	89.14
Unified	94.49	87.24	87.93	89.88

the discrepancies between the prediction scores and the corresponding labels. In general, the discrete perturbation $\bar{\epsilon}$ at the (u, v) -th location of the mask is computed as follows:

$$\bar{\epsilon}_{i(u,v)} = \begin{cases} \epsilon, & \text{if } \text{sign}\left(\frac{\partial \mathcal{L}_{CE}(\bar{x}_i')}{\partial \bar{M}_{i(u,v)}}\right) > 0 \\ -\epsilon, & \text{otherwise} \end{cases}, \quad (1)$$

where ϵ is the magnitude of perturbation. However, as seen on the left of Fig. 4, these perturbations lead to noise-like points within the forgery traces, which make the refined samples different from the real-world forgery samples. Since we consider both the adversarial property and the similarity with the real-world samples, the computation of perturbation is different from Eq. (1). The refined forgery added with the proposed unified perturbation is shown on the right of Fig. 4. Moreover, in Table 2, we compare the detection results of different perturbation settings. The results from different datasets demonstrate the importance of the similarity.

5 LIMITATION AND FUTURE WORK

The potential problem lies in the additional computational costs incurred by the refinement strategy. Therefore, we will mitigate these costs to achieve more efficient refinement in our future work. After that, for each forgery sample, we can perform multiple iterative refinements to further improve the forgery perception.

REFERENCES

- [1] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7132–7141.
- [2] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. 2020. Deepforensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2889–2898.