LAA-Net: Localized Artifact Attention Network for Quality-Agnostic and Generalizable Deepfake Detection

— Supplementary Material —

Overview

This document provides supplementary material complementing the main manuscript. It is structured as follows. First, the computation of the self-consistency loss and the ground truth generation of heatmaps are described. Second, more quantitative and qualitative results are provided. In particular, additional metrics are reported for both in-dataset and cross-dataset settings. Moreover, qualitative results comparing E-FPN and FPN are shown.

1. Self-Consistency Loss

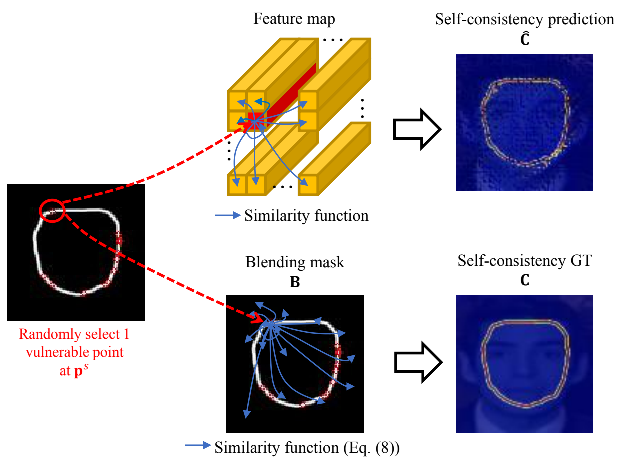

Figure 1. In order to generate the consistency map prediction $\hat{\mathbf{C}}$ as well as the associated ground truth $\mathbf{C}$, we first randomly select a vulnerable point located at $\mathbf{p}^s$. For computing $\hat{\mathbf{C}}$, we measure the similarity between the feature at $\mathbf{p}^s$ (red block) and the features generated from every point. Namely, we use the similarity function in [20]. As for $\mathbf{C}$, we measure the consistency values between the pixel at the $\mathbf{p}^s$ and all pixels in $\mathbf{B}$, as also described in Eq. (7) of the manuscript.

To clarify the calculation of the self-consistency loss, we show Figure 1, which illustrates the generation process of the predicted and the ground-truth, $\hat{\mathbf{C}}$ and $\mathbf{C}$, respectively. The self-consistency loss is a binary cross entropy loss between $\hat{\mathbf{C}}$ and $\mathbf{C}$.

2. Ground Truth Generation of Heatmaps

In this section, we provide more details regarding the generation of ground-truth heatmaps, described in Section 3.1.2. Firstly, a k-th vulnerable point, denoted as $\mathbf{p}^k$, is selected, as shown in Figure 3 (i). Secondly, we measure the height and the width of the blending mask $\mathbf{B}$ at the point $\mathbf{p}^k$ shown as orange lines in Figure 3 (ii). Using the calculated distances, a virtual bounding box is created, indicated by the blue box in Figure 3 (iii). Then, we identify overlapping boxes, illustrated by dashed-line green boxes in Figure 3 (iv), with the Intersection over Union (IoU) greater than a threshold ($t = 0.7$) compared to the virtual bounding box. A radius r_k (solid purple line in Figure 3 (v)) is calculated by forming a tight circle encompassing all these boxes. Finally, an *Unnormalized Gaussian Distribution*, shown as a red circle in Figure 3 (vi), is generated with a standard deviation $\sigma_k = \frac{1}{3} r_k$ (Eq. (4) of the manuscript). The steps are repeated for every vulnerable point $k \in [\![1, \text{card}(\mathcal{P})]\!]$. The final $\mathbf{H}$ is the superimposition of all g_{ij}^k.

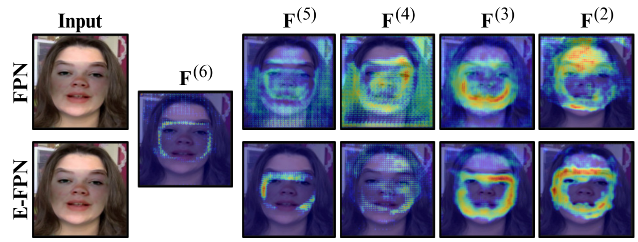

Figure 2. Feature visualization by gradCAM [13] between *E-FPN* and *FPN* with different integration of multi-scale layers. It shows that E-FPN can focus better on artifacts as compared to FPN. The setup details are provided in Table 4 as shown in the manuscript.

Method	Training Set		FF++ [12]				
	Real	Fake	ACC	AUC	AP	AR	mF1
Ours w/ BI [9]	✓		99.03	99.95	99.99	99.21	99.60
Ours w/ SBI [14]	✓		99.04	99.96	99.99	99.29	99.64

Table 1. In-dataset evaluation on FF++ [12] reported by ACC, AUC, AP, AR, and mF1.

3. Additional Results

In addition to AUC, we provide results using additional metrics, namely, Average Precision (AP), Average Recall (AR), Accuracy (ACC), and mean F1-score (mF1).

Table 1 and Table 2 report the results under the in-dataset and the cross-dataset settings, respectively. Overall, it can be seen that LAA-Net achieves better performances than other state-of-the-art methods.

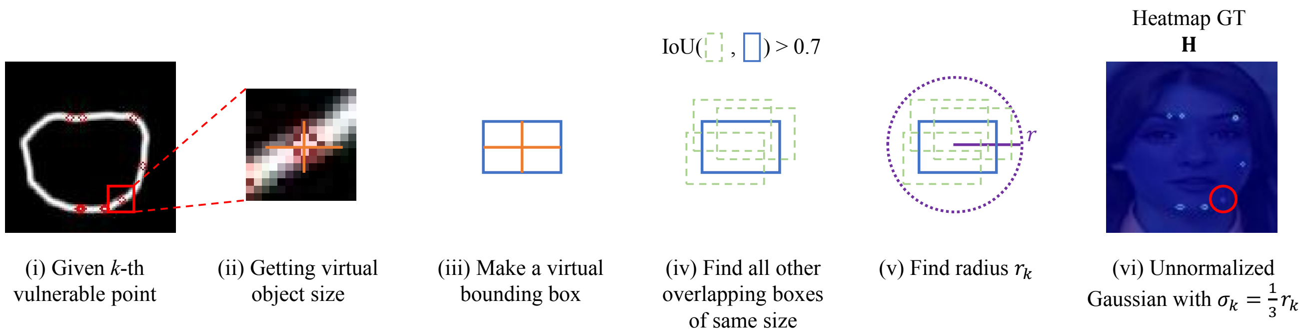

Figure 3. The generation process of ground truth heatmaps by producing using an *Unnormalized Gaussian Distribution* given a selected vulnerable point.

Method	Fake	Test set (%)															
		CDF2				DFW				DFD				DFDC			
		AUC	AP	AR	mF1	AUC	AP	AR	mF1	AUC	AP	AR	mF1	AUC	AP	AR	mF1
Xception [12]	✓	61.18	66.93	52.40	58.78	65.29	55.37	57.99	56.65	89.75	85.48	79.34	82.29	69.90	91.98	67.07	77.57
FaceXRay+BI [9]	✓	79.5	-	-	-	-	-	-	-	95.40	93.34	-	-	65.5	-	-	-
LRNet [15]	✓	53.20	-	-	-	-	-	-	-	52.29	-	-	-	-	-	-	-
LocalRL [3]	✓	78.26	-	-	-	-	-	-	-	89.24	-	-	-	76.53	-	-	-
TI^{2}Net [11]	✓	68.22	-	-	-	-	-	-	-	72.03	-	-	-	-	-	-	-
Multi-attentional [21]	✓	68.26	75.25	52.40	61.78	73.56	73.79	63.38	68.19	92.95	96.51	60.76	74.57	63.02	-	-	-
RECCE [2]	✓	70.93	70.35	59.48	64.46	68.16	54.41	56.59	55.48	98.26	79.42	69.57	74.17	-	-	-	-
SFDG [17]	✓	75.83	-	-	-	69.27	-	-	-	88.00	-	-	-	73.63	-	-	-
EIC+IIE [8]	✓	83.80	-	-	-	-	-	-	-	93.92	-	-	-	81.23	-	-	-
AltFreezing [18]	✓	89.50	-	-	-	-	-	-	-	98.50	-	-	-	-	-	-	-
CADDM [5]	✓	93.88	91.12	77.00	83.46	74.48	75.23	65.26	69.89	99.03	99.59	82.17	90.04	-	-	-	-
UCF [19]	✓	82.4	-	-	-	-	-	-	-	94.5	-	-	-	80.5	-	-	-
Controllable GS [7]	✓	84.97	-	-	-	-	-	-	-	-	-	-	-	81.65	-	-	-
PCL+I2G [20]		90.03	-	-	-	-	-	-	-	99.07	-	-	-	74.27	-	-	-
SBI [14]		93.18	85.16	82.68	83.90	67.47	55.87	55.82	55.85	97.56	92.79	89.49	91.11	86.15	93.24	71.58	80.99
AUNet [1]		92.77	-	-	-	-	-	-	-	99.22	-	-	-	86.16	-	-	-
Ours (w/ BI)		86.28	91.93	50.01	64.78	57.13	56.89	50.12	53.29	**99.51**	**99.80**	**95.47**	**97.59**	69.69	93.67	50.12	65.30
Ours (w/ SBI)		**95.40**	**97.64**	**87.71**	**92.41**	**80.03**	**81.08**	**65.66**	**72.56**	98.43	99.40	88.55	93.64	**86.94**	**97.70**	**73.37**	**83.81**

Table 2. Cross-dataset evaluation in terms of AUC, AP, AR, and mF1 (%) on CDF2 [10], DFW [22], DFD [6], and DFDC [4]. **Bold** and underlined highlight the best and the second-best performance, respectively. ✓symbol is used to depict methods that utilized both Real data and Fake data for training.

3.1. Qualitative Results: E-FPN versus FPN

A qualitative comparison between the proposed E-FPN and the traditional FPN with different fusion settings is reported in Figure 2. Using EfficientNet-B4 [16] (EFNB4) as our backbone, the $\mathbf{F}^{(6)}$ refers to the features extracted from the last convolution block in the backbone. In other words, this means that no FPN design is integrated. By gradually aggregating features from lower to higher resolution layers, we can observe the improvement of the forgery localization ability for both E-FPN and FPN. More notably, E-FPN produces more precise activations on the blending boundaries as compared to FPN. This can be explained by the fact that the E-FPN integrates a filtering mechanism for learning less noise. In contrast, FPN seems to consider regions outside the blending boundary, which results in lower performance as previously shown in Table 4 - Section 4.4 of the main manuscript.

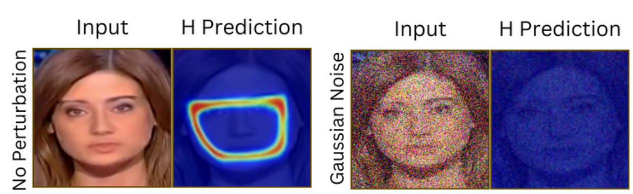

Figure 4. Detection of vulnerable points w/o and w/ Gaussian noise.

3.2. Qualitative Results: Gaussian Noise

In Table 2 of the main manuscript, the performance of LAA-Net declined significantly when encountering Gaussian Noise perturbations. One possible reason is that the introduction of noise elevates the difficulty of detecting the vulnerable points. To confirm that, we report the inference of the heatmap before and after applying a Gaussian Noise on a facial image in Figure 4. As it can be observed, the detection of vulnerable points is highly impacted with the introduction of a Gaussian noise.

3.3. Robustness to Compression

To assess the robustness of LAA-Net to compression, we test LAA-Net on the c23 version of FF++, and the overall AUC is equal to 89.30%.

References

[1] Weiming Bai, Yufan Liu, Zhipeng Zhang, Bing Li, and Weiming Hu. Aunet: Learning relations between action units for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24709–24719, 2023. 2

[2] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4103–4112, 2022. 2

[3] Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, Jilin Li, and R. Ji. Local relation learning for face forgery detection. In *AAAI Conference on Artificial Intelligence*, 2021. 2

[4] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton-Ferrer. The deepfake detection challenge (DFDC) preview dataset. *CoRR*, abs/1910.08854, 2019. 2

[5] Shichao Dong, Jin Wang, Renhe Ji, Jiajun Liang, Haoqiang Fan, and Zheng Ge. Implicit identity leakage: The stumbling block to improving deepfake detection generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3994–4004, 2023. 2

[6] Nick Dufour and Andrew Gully. Contributing data to deepfake detection research. https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html, 2019. 2

[7] Ying Guo, Cheng Zhen, and Pengfei Yan. Controllable guide-space for generalizable face forgery detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20818–20827, 2023. 2

[8] Baojin Huang, Zhongyuan Wang, Jifan Yang, Jiaxin Ai, Qin Zou, Qian Wang, and Dengpan Ye. Implicit identity driven deepfake face swapping detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4490–4499, 2023. 2

[9] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2

[10] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A new dataset for deepfake forensics. *CoRR*, abs/1909.12962, 2019. 2

[11] Baoping Liu, Bo Liu, Ming Ding, Tianqing Zhu, and Xin Yu. Ti2net: Temporal identity inconsistency network for deepfake detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4691–4700, 2023. 2

[12] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *International Conference on Computer Vision (ICCV)*, 2019. 1, 2

[13] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, abs/1610.02391, 2016. 1

[14] Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18720–18729, 2022. 1, 2

[15] Zekun Sun, Yujie Han, Zeyu Hua, Na Ruan, and Weijia Jia. Improving the efficiency and robustness of deepfakes detection through precise geometric features. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3608–3617, 2021. 2

[16] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *CoRR*, abs/1905.11946, 2019. 2

[17] Yuan Wang, Kun Yu, Chen Chen, Xiyuan Hu, and Silong Peng. Dynamic graph learning with content-guided spatial-frequency relation reasoning for deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7278–7287, 2023. 2

[18] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, and Houqiang Li. Altfreezing for more general video face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4129–4138, 2023. 2

[19] Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. Ucf: Uncovering common features for generalizable deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22412–22423, 2023. 2

[20] Eric Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning self-consistency for deepfake detection. In *ICCV 2021*, 2021. 1, 2

[21] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2185–2194, 2021. 2

[22] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. Wilddeepfake: A challenging real-world dataset for deepfake detection. *Proceedings of the 28th ACM International Conference on Multimedia*, 2020. 2