Supplementary Materials for An Information Theoretic Approach for Attention-Driven Face Forgery Detection

1 Results on DFDC

To further demonstrate the effectiveness of our proposed SIA, we conduct quantitative results on the DFDC dataset [3]. DFDC is a large-scale deepfake datasets that contains 1133 real videos and 4080 fake videos with several manipulated methods. The results in Tab.1 show that our method achieves SOTA performance compared with recently face forgery detection methods. Specifically, our method outperforms the Multi-Attentional method by around 3% in terms of ACC, which demonstrate the self-information can provide more guidance for attention mechanism and the effectiveness of the design of dual attention scheme.

Table 1. Performance on DFDC datasets in terms of ACC and AUC

Method	*DFDC*	
	ACC	AUC
Xception [2]	80.23	89.50
EfficientNet-b4 [16]	80.91	89.91
Add-Net [22]	79.90	89.85
RFM [17]	80.83	89.75
F3-Net [11]	77.66	88.39
MAT [21]	78.43	90.12
Ours	**81.31**	**90.96**

2 Robustness Analysis

In this section, we explore the robustness of our SIA module. Fig. 1 shows the quantitative results of baseline (EfficientNet-b4) and our method under different noises. Specifically, we conduct Gaussian blur with 3×3 kernel size, Gaussian noise, Salt and Pepper noise and jpeg compression on the test image of FF++ (HQ) dataset. The results show that compared with baseline, our SIA module is more robust especially under the Gaussian and Salt Pepper noise, which achieves about 6% on average improvement. This is because our SIA module help to adaptively extract more informative regions and further avoid the influence of invalid noise on input images.

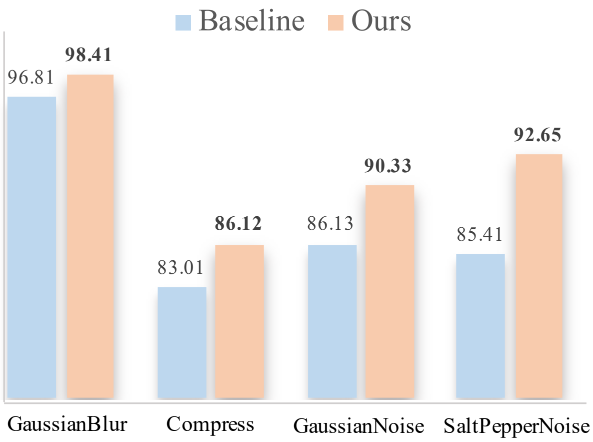

Fig. 1. Histograms of baseline and our method under different noises on FF (HQ) in terms of AUC. (Best viewed in color.)

3 Self-information Metric with Different Backbones

In this section, we first explore the impact of the backbones with different convolutional operations via several toy experiments. Then we introduce the self-information metric to explain the reason and further prove the effectiveness of the self-information metric in deepfake detection task.

According to [12,7], depthwise separable convolution based network, such as XceptionNet [2] and EfficientNet [16] is becoming the basic backbones in the deepfake detection tasks. To further explore the impact of the backbones and eliminate the influence of structures, three toy networks named *Norm network*, *CDC network* [20] and *Depthwise network* are constructed and detailed in Tab 2. The structure of these toy networks is the same while the convolutional operation is different. Specifically, *Norm network* means all the convolution method is vanilla convolution; *CDC network* replaces all the vanilla convolution as Central Difference Convolution which can capture the detailed features and prove effective in face anti-spoofing; *Depthwise network* replaces all the vanilla convolution as depthwise separable convolution.

Then we evaluate these networks on FaceFornsics++ dataset, the results are shown in Tab 3. We can observe that the performance of the *Norm network* drop significantly compared with both the *CDC network* and *Depthwise network* on both datasets while the parameters of the Norm network are far more than Depthwise network. This is very different from the experience of general classification tasks.

To prove the self-information metric really satisfy the deepfake detection task and explain the reason for this phenomenon at the same time, we calculate each

Table 2. Our toy network about different convolution methods, Conv (input, output, kernel_size, stride, group) and CDC(input, output, kernel_size, stride, group) denote the vanilla and central difference convolution with their corresponding parameters. For example, Conv (4,128,3,2,1) denotes input channel is 64, output channel is 128 with 3 × 3 kernel_size, stride is 2 and group is 1.

Layer	Norm network	CDC network	Depthwise network
Layer1	Conv(3,32,3,2,1) BatchNorm RELU	CDC(3,32,3,2,1) BatchNorm RELU	Conv(3,32,3,2,1) BatchNorm RELU

Layer2	Conv(32,64,3,1,1)
	BatchNorm
	RELU
		CDC(32,64,3,1,1)
		BatchNorm
		RELU
			Conv(32,32,3,2,32)
			Conv(32,64,1,1,1)
			BatchNorm
			RELU

Layer3	Conv(64,128,3,2,1)
	BatchNorm
	RELU
		CDC(64,128,3,2,1)
		BatchNorm
		RELU
			Conv(64,64,3,2,64)
			Conv(64,128,1,1,1)
			BatchNorm
			RELU

Layer4	Conv(128,256,3,1,1)
	BatchNorm
	RELU
		CDC(128,256,3,1,1)
		BatchNorm
		RELU
			Conv(128,128,3,1,128)
			Conv(128,256,1,1,1)
			BatchNorm
			RELU

Layer5	Conv(256,512,3,2,1)
	BatchNorm
	RELU
		CDC(256,512,3,2,1)
		BatchNorm
		RELU
			Conv(256,256,3,1,256)
			Conv(256,512,1,1,1)
			BatchNorm
			RELU

Layer6	Conv(512,1024,3,1,1)
	BatchNorm
	RELU
		CDC(512,1024,3,1,1)
		BatchNorm
		RELU
			Conv(512,512,3,1,512)
			Conv(512,1024,1,1,1)
			BatchNorm
			RELU

FC	pooling fc	pooling fc	pooling fc

Table 3. ACC(%) of three different toy models on DeepFakes HQ/LQ (DF HQ/LQ), Face2Face HQ/LQ (F2F HQ/LQ), FaceSwap HQ/LQ (FS HQ/LQ), NeuralTexures HQ/LQ (NT HQ/LQ) and Celeb-DF datasets. The last column represents the amount of parameters.

Toy-Model	DF HQ	DF LQ	F2F HQ	F2F LQ	FS HQ	FS LQ	NT HQ	NT LQ	Celeb-DF	Param
Norm	84.16	74.89	61.09	57.14	64.58	61.45	59.30	53.20	72.69	6.29M
CDC	90.05	78.95	82.90	64.69	89.87	67.74	81.93	62.14	89.65	6.29M
Depthwise	**91.38**	**86.99**	**91.47**	**77.11**	**90.19**	**79.83**	**82.97**	**66.59**	**95.43**	**4.90M**

layer's channel-wise average self-information of *Norm network*, *CDC network*, and *Depthwise network* on the high-quality of FaceFornsics++ dataset. Specifically, given t-th layer feature map f_k^t, the channel-wise self-information metric is defined as:

$$I_{avg}^t = \sum_{k=1}^{C} \frac{\sum_{i=1}^{H} \sum_{j=1}^{W} I(f_k^t(i,j))}{H \times W}, \tag{1}$$

where the H, W is the size of the feature map and C represents the number of channels I. The calculation method of I follows the original paper in E.q. (3).

The overall results are shown in Tab. 4, we can observe that the self-information metric of *Depthwise network* is larger than the *Norm network* and the *CDC network*, which means the depthwise separable convolutional based network can capture higher information content. Furthermore, by comparing Tab. 3 and Tab. 4, we find that the more self-information the feature map contained, the better performance the model has. So the self-information metric is satisfied with the deepfake detection task, which explains the rationality of our method from another perspective.

Table 4. Channel-wise average self-information of different toy models with their corresponding layers on the HQ of the FaceFornsics++ database.

Model	Layer2	Layer3	Layer4	Layer5	AVG
Norm Network	12.56	12.39	12.71	12.17	12.45
CDC Network	13.01	13.31	13.95	13.21	13.37
Depthwise Network	**13.56**	**13.97**	**14.69**	**13.88**	**14.03**

4 Verify the compatibility of SIA

To demonstrate the compatibility of SIA, we reproduce the recent SPSL [10], F3-Net [11] and DCL [15] and inject our SIA into their backbones, respectively. The results on DFDC and WildDeepfake are reported in Tab 5. We can observe that the performance achieve consistent improvement over three competitors after integrating our SIA module, which mainly benefits from the more efficient feature extractor based on self-information metric.

5 Related Work of Attention Mechanism

Attention mechanisms have been widely applied in many vision tasks [8,9,13]. The existing attention mechanisms can be categorized into channel-wise and spatial-wise. For channel attention, Hu *et al.* [5] first use squeeze and excitation operation to exploit the inter-channel relationship. For spatial attention, the work [18] introduces non-local operation in spacetime for images and videos to capture long-range context information. The work [4,19] extract informative features by blending cross-channel and spatial information together. What's more, some methods [1] improve the localization accuracy of CNN using an attention-based dropout layer. Other method [6] combine attention with multiple kernel selection to further improve performance. Furthermore, attention mechanism has been used in face forgery detection task. Stehouwer*et al.* [14] first use attention map to locate the forgery region under the supervision of the ground

Table 5. Quantitative results on DFDC and WildDeepfake (WDF) in terms of ACC and AUC.

Method	DFDC		WDF	
	ACC	AUC	ACC	AUC
SPSL	78.49	89.52	76.14	84.40
SPSL+SIA	**80.75**	**90.69**	**78.93**	**88.35**
F3-Net	77.66	88.39	80.66	87.53
F3-Net+SIA	**79.01**	**89.11**	**81.35**	**89.29**
DCL	79.27	88.77	79.19	89.27
DCL+SIA	**80.84**	**90.00**	**81.32**	**90.33**

truth mask. Multi-attentional [14] produce multiple spatial attention adaptively to make the network cover the local parts. Different from them whose generate attention map without any guidance and only consider the spatial-wise dimension, our SIA module provide a more suitable metric for both spatial-wise and channel-wise attention, which achieve better performance.

6 Future Work

In addition to the face forgery detection, we believe the proposed SIA can benefit the vision tasks that depend on extracting imperceptible but abnormal clues such as camouflaged object segmentation (COS), face anti-spoofing and image manipulation detection. Specifically, those tasks all suffer from the ignorance of the subtle features by deep models, e.g. camouflage object boundaries in COS, abnormal noises in face attacks and inconsistent image patterns in manipulated images. To this end, our proposed SIA module can be used to highlight those high-informative artifacts in both spatial and channel dimensions and preserve them through SI aggregation operation. Currently, we only evaluate our SIA module on the RGB domain. In future work, we will evaluate it in the frequency domain to further demonstrate its effectiveness and generality.

References

1. Choe, J., Shim, H.: Attention-based dropout layer for weakly supervised object localization. In: CVPR. pp. 2219–2228 (2019) 4
2. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: CVPR. pp. 1251–1258 (2017) 1, 2
3. Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., Ferrer, C.C.: The deepfake detection challenge dataset. arXiv preprint arXiv:2006.07397 (2020) 1

4. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: CVPR. pp. 3146–3154 (2019) 4
5. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR. pp. 7132–7141 (2018) 4
6. Li, X., Wang, W., Hu, X., Yang, J.: Selective kernel networks. In: CVPR. pp. 510–519 (2019) 4
7. Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-df: A new dataset for deepfake forensics. arXiv preprint arXiv:1909.12962 (2019) 2
8. Lin, G., Shen, C., Van Den Hengel, A., Reid, I.: Efficient piecewise training of deep structured models for semantic segmentation. In: CVPR. pp. 3194–3203 (2016) 4
9. Lin, Z., Feng, M., Santos, C.N.d., Yu, M., Xiang, B., Zhou, B., Bengio, Y.: A structured self-attentive sentence embedding. arXiv preprint arXiv:1703.03130 (2017) 4
10. Liu, H., Li, X., Zhou, W., Chen, Y., He, Y., Xue, H., Zhang, W., Yu, N.: Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In: CVPR. pp. 772–781 (2021) 4
11. Qian, Y., Yin, G., Sheng, L., Chen, Z., Shao, J.: Thinking in frequency: Face forgery detection by mining frequency-aware clues. In: ECCV. pp. 86–103. Springer (2020) 1, 4
12. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Face-forensics++: Learning to detect manipulated facial images. In: ICCV. pp. 1–11 (2019) 2
13. Shen, T., Zhou, T., Long, G., Jiang, J., Pan, S., Zhang, C.: Disan: Directional self-attention network for rnn/cnn-free language understanding. In: AAAI (2018) 4
14. Stehouwer, J., Dang, H., Liu, F., Liu, X., Jain, A.: On the detection of digital face manipulation. arXiv preprint arXiv:1910.01717 (2019) 4, 5
15. Sun, K., Yao, T., Chen, S., Ding, S., Li, J., Ji, R.: Dual contrastive learning for general face forgery detection. In: AAAI. vol. 36, pp. 2316–2324 (2022) 4
16. Tan, M., Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks. ICML (2019) 1, 2
17. Wang, C., Deng, W.: Representative forgery mining for fake face detection. In: CVPR. pp. 14923–14932 (2021) 1
18. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR. pp. 7794–7803 (2018) 4
19. Woo, S., Park, J., Lee, J.Y., So Kweon, I.: Cbam: Convolutional block attention module. In: ECCV. pp. 3–19 (2018) 4
20. Yu, Z., Zhao, C., Wang, Z., Qin, Y., Su, Z., Li, X., Zhou, F., Zhao, G.: Searching central difference convolutional networks for face anti-spoofing. In: CVPR. pp. 5295–5305 (2020) 2
21. Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., Yu, N.: Multi-attentional deepfake detection. arXiv preprint arXiv:2103.02406 (2021) 1
22. Zi, B., Chang, M., Chen, J., Ma, X., Jiang, Y.G.: Wilddeepfake: A challenging real-world dataset for deepfake detection. In: ACM MM. pp. 2382–2390 (2020) 1