Supplementary Material: On the Detection of Digital Face Manipulation

Hao Dang* Feng Liu * Joel Stehouwer * Xiaoming Liu Anil Jain
Department of Computer Science and Engineering
Michigan State University, East Lansing MI 48824

In this supplementary material, we provide some details and additional experimental results.

1. Details

Here we will further detail the Diverse Fake Face Dataset and proposed attention map, and analyze additional experiments.

1.1. DFFD Dataset Details

The DFFD was constructed from large and commonly used facial recognition datasets. This widespread use of FFHQ and CelebA validate our decision to utilize these as our real images, and the generation of manipulated images from them. As shown in Fig. 1, the DFFD encompasses large variance in both face size and human age, for both real and manipulated images. Details about the images from datasets used to construct the DFFD are available in Tab. 1.

1.2. Network Architecture Details

In Fig. 2, we show a simplified diagram of the placement of the attention layer within the Xception network. Due to its modularity, the attention layer can easily be added to any network, in a similar fashion to placing the attention layer in a different location in the Xception network.

2. Additional Experimental Results

2.1. Human Study

We conduct a human study to determine the ability of humans to distinguish between real and manipulated images in the DFFD. 10 humans participated in the study. This was accomplished using a random set of 110 images from the DFFD, where 10 images were taken from each row in Tab 1. For each image, the human was required to classify between Real, Entire Fake, and Partial Fake, and additionally required to provide polygon-based regions of interest (attention maps) for Partial Fakes. The results of this study are shown in Tab. 2. It is clear that humans have significant difficulty in the binary classification task (Entire Fake and Partial Fake are considered a single class), while our attention based solution performs almost perfectly.

*denotes equal contribution by the authors.

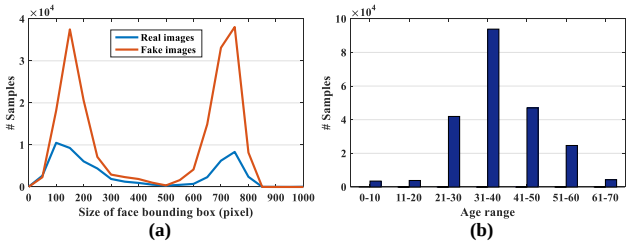

Figure 1. (a) Distribution of the face bounding box sizes (pixel) and (b) Age distribution of our DFFD.

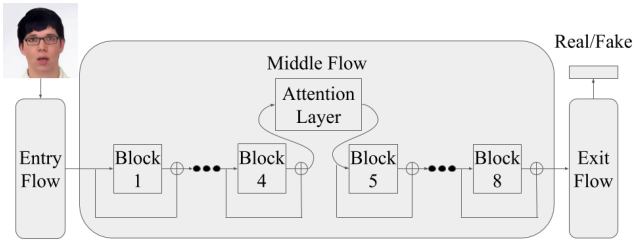

Figure 2. The overall architecture of XceptionNet and its enhancement with our proposed attention later. The original XceptionNet has entry flow, middle flow, and exit flow, where the middle flow is composed of 8 blocks. Our attention layer can be added after any of the blocks.

In Fig. 3, we show the manipulation maps produced by our proposed solution compared to the maps produced by humans. Humans focus largely on semantic concepts such as image quality, large artifacts, or strange lighting/color when judging between real and fake images. Due to this, humans do not detect the very subtle difference in the image "fingerprint", which our proposed solution is able to detect.

2.2. Additional Performance Evaluation

For our best performing model, Xception Regression Map with supervision, we conduct analysis in two aspects. *i*) Fig. 4 shows the worst 3 test samples among the real test faces and each fake types. For example, the images in the first column have the lowest Softmax probability of be-

Table 1. Statistics of our DFFD composition and protocol.

Dataset			# Total Samples	# Training	# Validation	# Testing	Average face width (pixel)
Real		FFHQ [4]	70,000	10,000	999	9,000	750
		CelebA [5]	202,599	9,974	997	8,979	200
		Original @ FaceForensics++ [6]	509,128	10,230	998	7,526	200
Fake	Id. Swap	DFL	49,235	10,006	1,007	38,222	200
		Deepfakes @ FaceForensics++ [6]	509,128	10,230	999	7,517	200
		FaceSwap @ FaceForensics++ [6]	406,140	8,123	770	6,056	200
	Exp. Swap	Face2Face @ FaceForensics++ [6]	509,128	10,212	970	7,554	200
	Attr. Manip.	FaceAPP [1]	18,416	6,000	1,000	5,000	700
		StarGAN [2]	79,960	10,000	1,000	35,960	150
	Entire Syn.	PGGAN [3]	200,000	19,957	1,998	17,950	750
		StyleGAN [4]	100,000	19,999	2,000	17,997	750

Table 2. Comparison between the proposed solution and humans for detecting manipulated images and localization of the manipulated regions. Larger values are better for all but the EER.

Method	ACC	AUC	EER	$TDR_{0.01\%}$	$TDR_{0.1\%}$	PBCA
Human	68.18	81.71	30.00	42.50	42.50	58.20
XceptionRegSup	**97.27**	**99.29**	**3.75**	**85.00**	**85.00**	**90.93**

Table 3. Fake face detection performance of the Xception Regression Map with supervision for each fake type.

Fake Type	AUC	EER	$TDR_{0.01\%}$	$TDR_{0.1\%}$
ID Manip.	99.43	3.11	65.16	77.76
EXP Manip.	99.40	3.40	71.23	80.87
Attr. Manip.	99.92	1.09	81.32	90.93
Entire Syn.	100.00	0.05	99.89	99.96

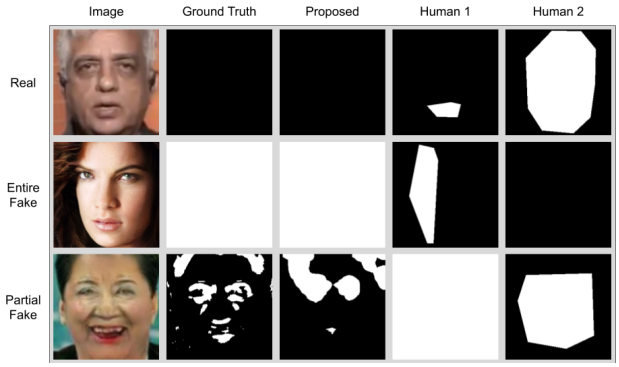

Figure 3. The attention maps produced by our proposed solution and humans during the human study.

ing the real class. Among these samples, some have heavy makeup, and others are of low image quality. Meanwhile, the failure cases for the manipulated or entirely synthetic images are high quality and devoid of defects or artifacts. *ii*) Tab. 3 shows the accuracy of testing samples in each fake type. The completely synthesized images are the easiest to detect. This is due to the artificial "fingerprint" these methods leave on the generated images, which is easily distinguishable from real images. In contrast, identity and expression manipulated images are the most challenging to detect, where image is of good quality and no noticeable artifacts exist, as in the 2nd and 3rd columns in Fig. 4.

Figure 4. Failure examples of the Xception with Regression Map under supervision. From left to right, the columns are top 3 worst samples of real, identity manipulated, expression manipulated, completely generated, and attribute modified, respectively.

2.3. Additional Ablation Study

In Fig 4, we show an ablation for the placement of the attention layer within the middle flow of the Xception network. Two trends emerge from this; *i*) the AUC and EER decrease as the attention layer is placed later in the network, and *ii*) the PBCA increases as the attention layer is placed later in the network. This second trend is expected, the network is able to produce a more finely-tuned attention map given more computational flexibility and depth. The first trend is more intriguing, because it shows that earlier focus from the attention map is more beneficial for the network than a finely-tuned attention map later. This earlier attention provides the network with additional time to inspect the features selected by the attention map in order to distinguish between real and manipulated images at a semantic level.

In Fig 5, we show the empirical decision for the threshold of 0.1 that we used to convert maps from continuous values (in the range [0,1]) to binary values. This provides strong performance in both graphs of Fig. 5, while being semantically reasonable. A modification of 0.1 corresponds to a modification of magnitude equal to 25 in the typical RGB range of [0,255]. While a modification of small magnitude

Table 4. The performance of the attention map at different placements in the middle flow of the XceptionNet architecture.

Map position	AUC	EER	$TDR_{0.01\%}$	$TDR_{0.1\%}$	PBCA
Block1	99.82	**1.69**	71.46	**92.80**	83.30
Block2	**99.84**	1.72	67.95	90.14	87.41
Block3	99.50	2.82	49.06	72.50	88.14
Block4	99.64	2.23	**83.83**	90.78	88.44
Block5	99.49	2.62	82.70	89.03	88.40
Block6	99.72	2.28	63.08	86.02	87.41
Block7	99.78	1.79	28.51	88.98	88.39
Block8	98.62	4.42	74.24	79.95	**88.96**

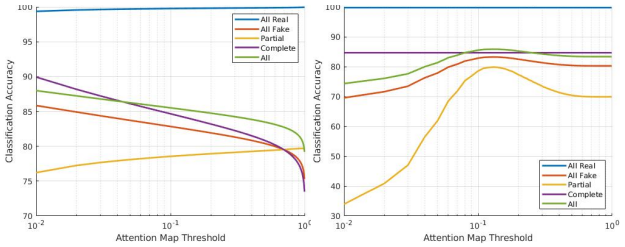

Figure 5. The attention map estimation performance of the proposed method when using different thresholds to binarize the predicted map (a) and the ground truth map (b). The threshold for the other map in either case was 0.1.

(< 10) is almost undetectable by a human, a modification of larger magnitude (> 25) is significant. Therefore, all experiments presented utilized this empirical threshold value of 0.1.

References

[1] FaceApp. https://faceapp.com/app. Accessed: 2019-09-04. 2

[2] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018. 2

[3] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 2

[4] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2

[5] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 2

[6] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *ICCV*, 2019. 2