CelebA-Spoof: Large-Scale Face Anti-Spoofing Dataset with Rich Annotations (Supplementary Material)

Anonymous ECCV submission

Paper ID 1485

1 Detail Information of CelebA-Spoof Dataset

Spoof Images in CelebA. As shown in Figure 1. In CelebA [6], there are 347 "spoof" images, including poster, advertisements and portrait *etc.* For spoof instruments selection and live data collection on CelebA-Spoof, we manually examine these images and remove them.

Table 1. Input sensor split in CelebA-Spoof, there are 24 different input sensors which are split into 3 groups based on image quality

	Sensor	Dataset	Pix. (MP)	Release		Sensor	Dataset	Pix. (MP)	Release		Sensor	Dataset	Pix. (MP)	Release
Low-Quality Sensor	Honor V8	train test val	1200	2016	Middle-Quality Sensor	vivo X20	train test val	1200	2018	High-Quality Sensor	HUAWEI P30	train test val	4000	2019
	OPPO R9	train test val	1300	2016		Gionee S11	train test val	1300	2018					
	HUAWEI MediaPad M5	train test	1200	2016		vivo Y85	train val	1600	2018					
	Xiaomi Mi Note3	train test val	1200	2016		Hisense H11	train val	2000	2018					
	Gionee S9	train test val	1300	2016		iphone XR	train	1200	2018		meizu 16S	train test val	4800	2019
	Logitech C670i	train	1200	2016		OPPO A5	train	1300	2018					
	ThinkPad T450	train	800	2016										
	Moto X4	train test val	1200	2017		OPPO R17	train	1600	2018					
	vivo X7	train test val	1200	2017		OPPO A3	train test val	1200	2019		vivo NEX 3	train	6400	2019
	Dell 5289	train	800	2017		Xiaomi 8	train test val	1200	2019					
	OPPO A73	train	1600	2017		vivo Y93	train test val	1300	2019					

Fig. 1. Representative examples of the "spoof " images in CelebA

Input Sensor Split. As shown in Table 1, according to imaging quality, we split 24 input sensors into 3 groups: *low-quality sensor*, *middle-quality sensor* and *high-quality sensor*. In detail, an input sensor is not necessarily used in the all training, verification and testing set, so we specify which dataset these input

sensors would cover. Specifically, for cross-domain benchmark in CelebA-Spoof, only input sensors which are both used in training set and testing set are selected.

2 Experimental Details

Formulations of Evaluation Metrics. To establish a comprehensive benchmark, we unify 7 commonly used metrics (*i.e.* APCER, BPCER, ACER, EER, HTER, AUC and FPR@Recall). Besides AUC, EER and FPR@Recall which are the most common metrics of classification tasks, we list definitions and formulations of other metrics. 1) *APCER, BPCER and ACER.* Refer to [1,5], Attack Presentation Classification Error Rate (APCER) is used to evaluate the classification performance of models for spoof images. Bona Fide Presentation Classification Error Rate (BPCER) is used to evaluate the classification performance of models for live images:

$$APCER_{\mathcal{S}^s} = \frac{1}{N_{\mathcal{S}^s}} \sum_{i=1}^{N_{\mathcal{S}^s}} (1 - Res_i) \tag{1}$$

$$APCER = \max(APCER_{\mathcal{S}^s{}_1}, APCER_{\mathcal{S}^s{}_2} ... APCER_{\mathcal{S}^s{}_k}) \tag{2}$$

$$BPCER_{\mathcal{S}^f} = \frac{1}{N_{\mathcal{S}^f}} \sum_{i=1}^{N_{\mathcal{S}^f}} Res_i \tag{3}$$

$$BPCER = \frac{1}{N_{liv.}} \sum_{i=1}^{N_{liv.}} Res_i \tag{4}$$

$$ACER = \frac{(APCER + BPCER)}{2} \tag{5}$$

where, $N_{\mathcal{S}^s}$ is the number of the spoof images of the given spoof type. $N_{\mathcal{S}^s}$ is the number of the live images of the given face attribute. $N_{liv.}$ is the number of all live images. Res_i takes the value 1 if the ith images is classified as an spoof image and 0 if classified as live image. $\text{APCER}_{\mathcal{S}^s}$ is computed separately for each micro-defined spoof type (*e.g.* "photo", "A4", "poster") and APCER is the highest $\text{APCER}_{\mathcal{S}^s}$ which represent the "worst case scenerio", k is the number of micro-defined spoof type. Specifically, in CelebA-Spoof, we define $\text{BPCER}_{\mathcal{S}^f}$ which is computed separately for each face attribute. To summarize the overall performance of live images and spoof images, the Average Classification Error Rate (ACER) is used, which is the average of the APCER and the BPCER at the decision threshold defined by the Equal Error Rate (EER) on the testing set. 2) *HTER.* The aforementioned metrics are employed on intra-dataset (CelebA-Spoof) evaluation. For cross-dataset evaluation, HTER [3] is used extensively:

$$HTER(D_2) = \frac{FAR(\tau(D1), (D_2)) + FRR(\tau(D1), (D_2))}{2} \tag{6}$$

Table 2. The mAP result of single-task and multi-task. There is huge space to improve the learning of $\mathcal{S}^{\text{i}}$ in multi-task fashion. **Bolds** are the best results

Attribute	Model	mAP (%)
$\mathcal{S}^{\text{s}}$	AENet$_{\mathcal{S}^{\text{s}}}$	45.7
	AENet$_{\mathcal{C},\mathcal{S}}$	**46.2**
$\mathcal{S}^{\text{f}}$	AENet$_{\mathcal{S}^{\text{f}}}$	68.5
	AENet$_{\mathcal{C},\mathcal{S}}$	**70.5**
$\mathcal{S}^{\text{i}}$	AENet$_{\mathcal{S}^{\text{i}}}$	**57.1**
	AENet$_{\mathcal{C},\mathcal{S}}$	43.3

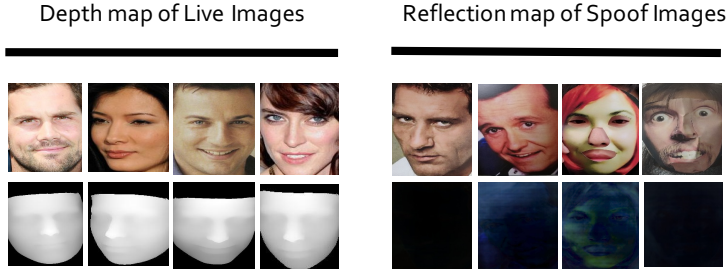

Fig. 2. All live images have depth maps, but only the second and the third spoof image has reflection artifacts. Zoom in for better visualization

where $\tau(D_n)$ is a threshold, D_n is the dataset, False Acceptance Rate (FAR) and False Rejection Rate (FRR) is the value in D_2. In cross-dataset evaluation, the value of $\tau(D_n)$ is estimated on the EER using the testing set of the dataset D_1. In this equation, when $D_1 \neq D_2$, we have the cross-dataset evaluation.

The Limitations of Reflection Map. For ablation study of geometric information, we do not use reflection maps as unique binary supervision. This is because only parts of spoof images show reflect artifacts as shown in Figure 2. In this figure, only the second and the third spoof image shows reflect artifacts, the reflection map for other spoof images is zero. However, each live image has its corresponding depth map.

Multi Task and Single Task. Besides ablation study of semantic information. We compare AENet$_{\mathcal{S}^{\text{f}}}$, AENet$_{\mathcal{S}^{\text{s}}}$ and AENet$_{\mathcal{S}^{\text{i}}}$ with AENet$_{\mathcal{C},\mathcal{S}}$ to explore whether multi-task learning can promote classification performance of these semantic information. In detail, as mentioned in model setting of Sec. *Ablation Study on CelebA-Spoof*. AENet$_{\mathcal{S}^{\text{f}}}$, AENet$_{\mathcal{S}^{\text{s}}}$ and AENet$_{\mathcal{S}^{\text{i}}}$ are trained for classification of each semantic information. As shown in the Table 2. It shows that the mAP performance of $\mathcal{S}^{\text{f}}$ and the $\mathcal{S}^{\text{s}}$ in AENet$_{\mathcal{C},\mathcal{S}}$ is better than AENet$_{\mathcal{S}^{\text{f}}}$ and AENet$_{\mathcal{S}^{\text{s}}}$. Specifically, these two semantic information are proven crucial to improve classification of live/spoof images in ablation study. Besides, the mAP performance of $\mathcal{S}^{\text{i}}$ of AENet$_{\mathcal{C},\mathcal{S}}$ is worse than AENet$_{\mathcal{S}^{\text{i}}}$. This is because we set the $\lambda = 0.01$ of $\mathcal{S}^{\text{i}}$ in the multi-task training but $\lambda = 1$ for all single task model. This small value let $\mathcal{S}^{\text{i}}$ difficult to converge in multi task learning.

Table 3. Intro-dataset Benchmark results of CelebA-Spoof. AENet$_{\mathcal{C},\mathcal{S},\mathcal{G}}$ achieved the best result. **Bolds** are the best results; ↑ means bigger value is better; ↓ means smaller value is better

Model	Parm. (MB)	Recall (%)↑			AUC↑	EER (%)↓	APCER (%)↓	BPCER (%)↓	ACER (%)↓
		FPR = 1%	FPR = 0.5%	FPR = 0.1%					
AENet$_{\mathcal{C},\mathcal{G}}$	79.9	98.3	97.2	91.4	**0.9982**	0.012	4.98	1.26	3.12
AENet$_{\mathcal{C},\mathcal{S}}$	79.9	98.5	97.8	**94.3**	0.9980	0.013	4.22	1.21	2.71
AENet$_{\mathcal{C},\mathcal{S},\mathcal{G}}$	79.9	**99.2**	**98.4**	94.2	0.9981	**0.009**	**3.72**	**0.82**	**2.27**

Table 4. Cross-domain benchmark results of CelebA-Spoof. **Bolds** are the best results; ↑ means bigger value is better; ↓ means smaller value is better

Protocol	Model	Recall (%) ↑			AUC↑	EER (%)↓	APCER (%)↓	BPCER (%)↓	ACER (%)↓
		FPR = 1%	FPR = 0.5%	FPR = 0.1%					
1	Baseline	94.6	92.3	**86.4**	0.985	0.038	9.19	3.84	6.515
	AENet$_{\mathcal{C},\mathcal{G}}$	93.7	89.7	73.1	0.984	0.034	7.66	3.11	5.39
	AENet$_{\mathcal{C},\mathcal{S}}$	96.5	**93.1**	83.4	0.992	0.023	3.78	1.8	2.79
	AENet$_{\mathcal{C},\mathcal{S},\mathcal{G}}$	**96.9**	93.0	83.5	**0.996**	**0.018**	**3.00**	**1.48**	**2.24**
2	Baseline	#	#	#	0.996±0.003	0.018±0.009	7.44±2.62	1.81±0.9	4.63±1.66
	AENet$_{\mathcal{C},\mathcal{G}}$	#	#	#	0.994±0.006	0.017±0.006	9.16±1.97	1.56±1.68	5.36±1.23
	AENet$_{\mathcal{C},\mathcal{S}}$	#	#	#	0.996±0.003	0.012±0.009	5.08±4.41	**0.95±0.68**	4.02±2.6
	AENet$_{\mathcal{C},\mathcal{S},\mathcal{G}}$	#	#	#	**0.997±0.003**	**0.013±0.012**	**4.77±4.12**	1.23±1.06	**3.00±2.9**

Table 5. Cross-dataset benchmark results of CelebA-Spoof. AENet$_{\mathcal{C},\mathcal{S},\mathcal{G}}$ achieves the best generalization performance. **Bolds** are the best results; ↑ means bigger value is better; ↓ means smaller value is better

Model	Training	Testing	HTER (%) ↓
Baseline	CelebA-Spoof	CASIA-MFSD	20.1
AENet$_{\mathcal{C},\mathcal{G}}$	CelebA-Spoof	CASIA-MFSD	18.2
AENet$_{\mathcal{C},\mathcal{S}}$	CelebA-Spoof	CASIA-MFSD	17.7
AENet$_{\mathcal{C},\mathcal{S},\mathcal{G}}$	CelebA-Spoof	CASIA-MFSD	**13.1**

3 Benchmark on Heavier Model

In order to build a comprehensive benchmark, besides ResNet-18 [4], we also provide the corresponding results based on a heavier backbone, *i.e.* Xception [2]. All the results on the following 3 benchmarks are based on Xception. Detail information about benchmark based on ResNet-18 is shown in paper. 1) *Intra-Dataset Benchmark.* As shown in Table 3, AENet$_{\mathcal{C},\mathcal{S},\mathcal{G}}$ based on Xception achieve better performance comparing to AENet$_{\mathcal{C},\mathcal{S},\mathcal{G}}$ based on ResNet-18, especially when FPR is smaller (*i.e.* FPR=0.5% and FPR=0.1%). This is because model with heavier parameters can achieve better robustness. 2) *Cross-domain Benchmark.* As shown in Table 4, AENet$_{\mathcal{C},\mathcal{S},\mathcal{G}}$ based on Xception achieve the better performance than AENet$_{\mathcal{C},\mathcal{S},\mathcal{G}}$ based on ResNet-18. And in protocol 1, comparing to baseline based on Xception. AENet$_{\mathcal{C},\mathcal{S},\mathcal{G}}$ based on Xception outperforms baseline by 67.3% in APCER. 3) *Cross-dataset Benchmark.* As shown in Table 5. Performance of models based on Xception is worse than models based on ResNet-18. This is because models with heavier parameters tend to fit the training data.

References

1. Boulkenafet, Z., Komulainen, J., Li, L., Feng, X., Hadid, A.: Oulu-npu: A mobile face presentation attack database with real-world variations. In: FG. pp. 612–618. IEEE (2017) 2
2. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: CVPR. pp. 1251–1258 (2017) 4
3. de Freitas Pereira, T., Anjos, A., De Martino, J.M., Marcel, S.: Can face anti-spoofing countermeasures work in a real world scenario? In: ICB. pp. 1–8. IEEE (2013) 2
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016) 4
5. Liu, Y., Stehouwer, J., Jourabloo, A., Liu, X.: Deep tree learning for zero-shot face anti-spoofing. In: CVPR. pp. 4680–4689 (2019) 2
6. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: ICCV (2015) 1