SUPPLEMENTARY MATERIALS FOR GUIDED ADVERSARIAL WATERMARKS FOR PHYSICAL ATTACKS ON FACE RECOGNITION

Jiaxuan Zhu[1] *Siyu Xia*[1] *Ming Shao*[2]

[1] School of Automation, Southeast University, Nanjing, Jiangsu, 210096, China
[2]Computer and Information Science, University of Massachusetts, Dartmouth, MA 02747, USA

1. SUPPLEMENTARY EXPERIMENT

1.1. Results on different datasets

In Table 1, we discuss the efficacy of our method on different datasets for face verification. To that tend, Cross-Pose LFW (CPLFW) [1] dataset was introduced in the experiment, and the same metrics are used. This experiment reveals that our method is general and applicable to different datasets in terms of face verification adversarial attack.

1.2. Discussion on watermarks

In Table 2, we discuss the impacts of the scale factor s, and we find a larger s would benefit the proposed adversarial attack. In Table 3, we explore different watermarks and their efficacy in attacks. In general, all three watermarks apply to our method but show different performance [1] [2]. It shows that FR models have different preferences over the watermarks. For example, while "resnet50+center loss" prefers the Berkerly logo, "resnet50+arcface loss" works better with the MIT Logo. In Table 3, we discuss the impact of ϵ on our method and conclude that a larger ϵ could lead to a better attack performance. In Figure 1, we also discuss the transformation process of heat maps, which demonstrates how Eigen-CAM guides the embedded adversarial watermark. These results demonstrate that the transformation of heat maps resulting from adversarial attacks might not have a uniform pattern. We speculate that our attack method might mislead the FR models to extract key features from adversarial watermarks or other less critical areas instead of faces, thereby affecting face recognition results.

[1]These watermarks can be found at `https://github.com/jiaxiaojunQAQ/Adv-watermark.git`

Table 1: Evaluations on different datasets.

Dataset	Original Accuray(%)	TA/SRA(%)
LFW	96.39	71.62/53.14
CPLFW [1]	88.47	58.52/87.75

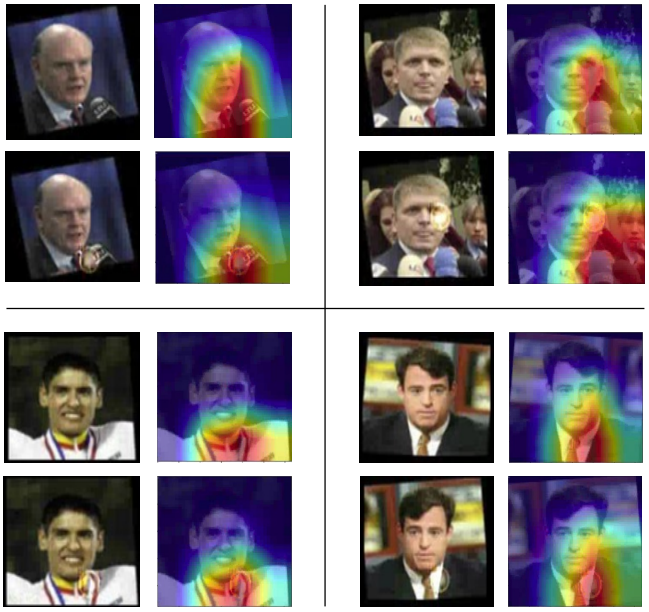

Fig. 1: Visualization of Eigen-CAM heat maps before (first row) and after (second row) adversarial attacks. The target model is "resnet50+center loss." Four examples are included in separate areas. In each area, the left image is the face image, while the right image is its heat map. The first row represents the heat map visualization of the original images, while the second row represents the heat map visualization of the adversarial samples.

1.3. Qualitative results

In Figure 2, four verification results on LFW test data using our method with a transparency of $\alpha = 0.5$ is shown. Note the first/second columns show the results of before/after attack, respectively. It demonstrates that our method could enlarge the distance between input face images and achieve the goal of privacy protection. Figure 3 and Figure 4 visualize adversarial samples and adversarial watermarks, respectively. These results demonstrate the influence of transparency parameter α on the proposed attack method and reveal the fact that em-

Table 2: Experimental results of the proposed method with different watermark sizes. Transparency parameter is set to $\alpha = 0.5$.

FR Models	Original Accuracy(%)	TA/SRA(%) s = 2/5	s = 1/3	s = 1/4	s = 1/5
resnet50+center loss *	96.39	60.53/76.49	66.05/64.91	71.62/53.14	86.25/23.18
resnet18+center loss	95.27	71.58/52.92	78.21/38.32	85.20/25.20	91.40/10.43
resnet50+cosface	99.73	82.08/35.40	86.51/26.49	89.22/21.25	99.48/0.70
resnet18+cosface	99.47	73.57/52.09	81.85/35.52	94.77/9.23	98.95/0.77
resnet50+arcface	96.07	61.22/74.64	66.96/62.34	72.23/50.85	85.68/22.69
resnet18+arcface	92.60	70.93/51.07	75.31/40.59	82.26/28.59	89.05/9.14
resnet50+sphereface	97.27	63.82/70.99	69.90/57.62	78.62/39.66	89.48/17.03
resnet18+sphereface	94.32	66.73/62.95	69.80/56.33	78.48/38.12	88.08/16.80

Table 3: Experimental results of our method with different watermarks and ϵ. When testing different watermarks, default parameters are used but three logos are evaluated. When testing different ϵ, the same default values are used but varying ϵ.

FR Models	Original Accuracy(%)	Test Accuracy (TA)/Success Rates of Attacking (SRA) (%) Berkeley Logo	MIT Logo	SU Logo	ϵ = 8/255	ϵ = 16/255	ϵ = 64/255
resnet50+center loss *	96.39	71.62/53.14	78.91/37.30	75.18/45.12	79.17/37.28	77.55/40.22	71.62/53.14
resnet18+center loss	95.27	85.20/25.20	83.25/28.39	83.52/27.80	89.35/15.45	88.17/17.87	85.20/25.20
resnet50+cosface	99.73	89.22/21.25	93.97/11.69	92.62/14.30	98.52/2.58	97.80/3.98	89.22/21.25
resnet18+cosface	99.47	94.77/9.23	93.63/11.52	92.55/13.76	97.75/3.59	97.08/4.70	94.77/9.23
resnet50+arcface	96.07	72.23/50.85	70.62/54.50	69.76/56.57	80.23/34.44	82.37/28.75	72.23/50.85
resnet18+arcface	92.60	82.26/28.59	81.73/29.72	80.33/32.24	85.75/15.38	85.53/18.89	82.26/28.59
resnet50+sphereface	97.27	78.62/39.66	74.62/48.32	74.50/48.85	83.92/28.08	82.25/31.60	78.62/39.66
resnet18+sphereface	94.32	78.48/38.12	76.77/42.06	74.48/46.25	84.78/23.90	83.95/25.94	78.48/38.12

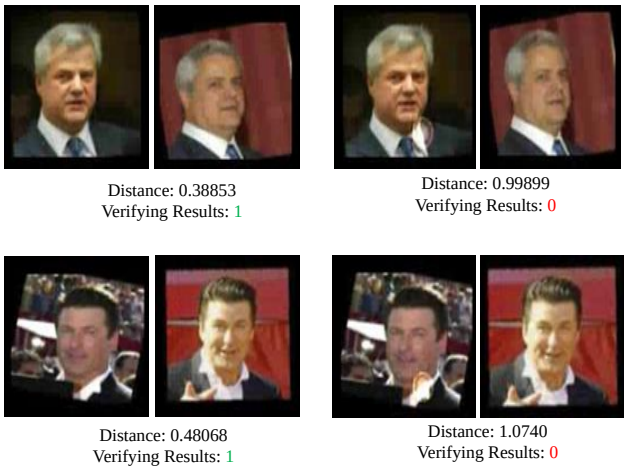

Fig. 2: Qualitative results of face verification before and after attack on the LFW dataset. We select "resnet50+center loss" as the target FR model and set the distance threshold $\theta = 0.8$.

bedded watermarks may occlude face images given a larger value of α. In addition, Figure 3 shows that Eigen-CAM based adversarial watermarks guidance may suggest varied locations. In general, lower α leads to minor occlusion on face images, while $\alpha \geqslant 0.5$ usually results in more effective attacks on FR models. To stabilize the attack while managing the occlusions, we set $\alpha = 0.5$ as the default.

Fig. 3: Comparison results of adversarial samples with different transparency parameters α on LFW dataset. The first row denotes original face images, and the second to sixth rows denote adversarial samples generated by our method using different transparency parameters α.

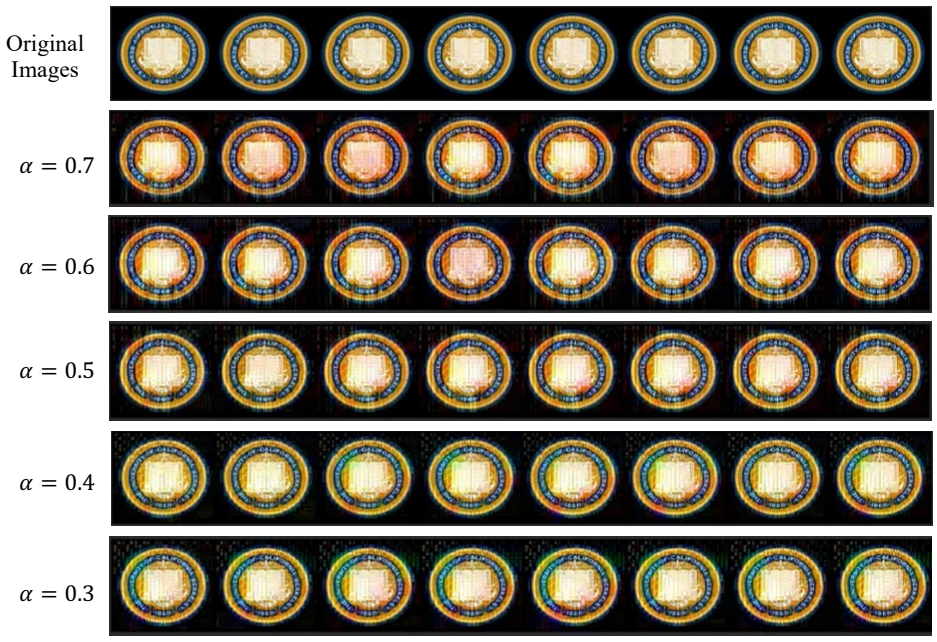

Fig. 4: Comparison results of adversarial watermarks with different transparency parameters α on the LFW dataset. The first row denotes original watermark images, and the second to sixth rows denote adversarial watermarks generated by our method using different transparency parameters α.

2. REFERENCES

[1] T. Zheng and W. Deng, "Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments," Tech. Rep. 18-01, Beijing University of Posts and Telecommunications, February 2018.

[2] Xiaojun Jia, Xingxing Wei, Xiaochun Cao, and Xiaoguang Han, "Adv-watermark: A novel watermark perturbation for adversarial examples," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1579–1587.