Supplementary Material for Improving Robustness of Facial Landmark Detection by Defending against Adversarial Attacks

Congcong Zhu, Xiaoqiang Li*, Jide Li, Songmin Dai
School of Computer Engineering and Science, Shanghai University, Shanghai, China
{ congcongzhu, xqli, iavtvai, laodar }@shu.edu.cn

Abstract

In this document, some qualitative results of adversarial examples are used to show different strategies of adversarial attacks. We also provide more details of hyper-parameters. Moreover, representative samples of the proposed Masked-300W Dataset are demonstrated.

1. Different strategies of adversarial attacks

In this work, the purpose of adversarial attacks is to disturb the detector's localization for facial landmarks. To this end, a conditional GAN (CGAN) is introduced as the attacker to generate adversarial perturbations, which exploits structure map S to guide the attacker to be aware of facial semantic regions. In this section, we compare the conditional GAN with a general GAN. Specifically, the conditional GAN is optimized by the following function:

$$\min_G \max_D V(D,G) = \mathbb{E}[\log D(I,S)] + \mathbb{E}[\log(1 - D(I + G(I,S,z),S))]. \quad (1)$$

The general GAN is optimized by the following function:

$$\min_G \max_D V(D,G) = \mathbb{E}[\log D(I)] + \mathbb{E}[\log(1 - D(I + G(I,z)))]. \quad (2)$$

Note that G denotes the attacker in our framework. Quantitative results are presented in Table 1. We observe that the CGAN can better improve SAAT than the general GAN.

Methods	Challenging	Common	Full
HG*1+SAAT with CGAN	5.10	2.96	3.38
HG*1+SAAT with GAN	5.18	3.03	3.45

Table 1. Quantitative results on 300W.

In our Ablation study, we report the performance of semantic reconstruction loss. In this section, we show qualitative results proving that semantic reconstruction plays an important role in the perturbation generation, see Figure 1. We can see that CGAN+reconstruction loss achieves the best generation performance. Using CGAN alone will produce false samples. This means that semantic reconstruction can significantly improve the stability of the attacker.

*Corresponding author.

2. Adaptive weight

Although our SAAT can generate high-quality perturbations, false samples may appear in the early stages of end-to-end training. To avoid the performance degeneration of the detector caused by false samples, we propose the adaptive weight β to adjust the contribution of adversarial examples to the detector. In the Ablation study of our paper, we have shown the effect of this adaptive weight. In this section, we will show some qualitative results, see Figure 1.

3. Masked-300W

To evaluate masked face alignment, we introduce a new masked face alignment dataset Masked-300W base on 300W [4], which contains the same test sets as 300W: Common set (LFPW and HELEN test sets, 554 images), Challenging set (IBUG set, 135 images), and Full set (the union of the former two sets, 689 images). Each face of this dataset is worn with a blue medical mask following Simulated Masked Face Recognition Dataset (SMFRD) [5]. To further improve the accuracy of the mask position, we directly use real landmarks to locate the position of each mask and warp it. As shown in Figure 2, these faces are severely occluded under unconstrained environments.

Additional qualitative results of the proposed SAAT approach are illustrated in Figure 3. We can see that the baseline cannot handle occluded faces because of lacking the shape constraint. The proposed SAAT can improve the robustness of the baseline against severe occlusion and partial observability.

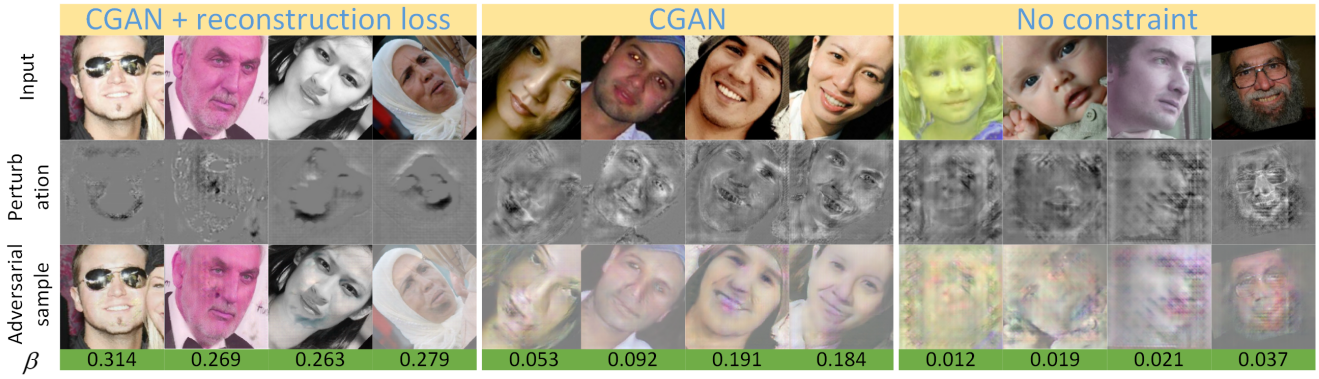

Figure 1. Qualitative results of different perturbation generation strategies.

Figure 2. Representative samples of the proposed Masked-300W Dataset.

4. Discussion

We think that SAAT is not a replacement but a complement for handcrafted transformations. Handcrafted transformations are still necessary and are applied in the training phase. However, the transformation is human prior based, which cannot adaptively explore the weaknesses of detectors against attacks from the real world. Researchers [1, 3, 2] have proven that exploring and remedying the weaknesses of models during training can improve the robustness. Therefore, we believe that the proposed SAAT can improve the robustness of detectors against attacks.

References

[1] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 2

[2] Alexey Kurakin, Ian Goodfellow, Samy Bengio, et al. Adversarial examples in the physical world. 2

[3] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *European symposium on security and privacy*, pages 372–387. IEEE, 2016. 2

[4] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *ICCVW*, pages 397–403, 2013. 1

[5] Zhongyuan Wang, Guangcheng Wang, Baojin Huang, Zhangyang Xiong, Qi Hong, Hao Wu, Peng Yi, Kui Jiang, Nanxi Wang, Yingjiao Pei, et al. Masked face recognition dataset and application. *arXiv preprint arXiv:2003.09093*, 2020. 1

Figure 3. Qualitative results on Masked-300W Dataset.