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

Many recent developments in facial landmark detection have been driven by stacking model parameters or augmenting annotations. However, three subsequent challenges remain, including 1) an increase in computational overhead, 2) the risk of overfitting caused by increasing model parameters, and 3) the burden of labor-intensive annotation by humans. We argue that exploring the weaknesses of the detector so as to remedy them is a promising method of robust facial landmark detection. To achieve this, we propose a sample-adaptive adversarial training (SAAT) approach to interactively optimize an attacker and a detector, which improves facial landmark detection as a defense against sample-adaptive black-box attacks. By leveraging adversarial attacks, the proposed SAAT exploits adversarial perturbations beyond the handcrafted transformations to improve the detector. Specifically, an attacker generates adversarial perturbations to reflect the weakness of the detector. Then, the detector must improve its robustness to adversarial perturbations to defend against adversarial attacks. Moreover, a sample-adaptive weight is designed to balance the risks and benefits of augmenting adversarial examples to train the detector. We also introduce a masked face alignment dataset, Masked-300W, to evaluate our method. Experiments show that our SAAT performed comparably to existing state-of-the-art methods. The dataset and model are publicly available at https://github.com/zhuccly/SAAT.

1. Introduction

Recently, facial landmark detection has been significantly improved by many works [36, 41, 42, 46, 51]. To achieve new breakthroughs, researchers proposed multi-stage stacked networks [3, 12, 25, 10, 24, 28, 40]. Methods such as [12, 25] use the two-stage architecture to regress the facial shape in a coarse-to-fine manner. In [35, 28], the multi-stage stacked hourglass networks (SHG) are employed to predict the landmark heatmaps. In [19, 24, 40], the additional sub-networks are equipped into the multi-stage stacked hourglass networks to improve the fitting performance further. Although the increasing number of model parameters leads to breakthroughs, the redundant parameters may raise the risk of these models overfitting on the training dataset, especially in small datasets that their data distributions are usually unbalanced [4, 31]. Even the data distribution of existing large-scale datasets may be unbalanced. For example, 300VW dataset [33] contains 95,192 training frames. However, these frames are only collected from 50 videos containing 44 different persons under 34 scenes. Because this dataset is large but not diverse, multi-stage stacked models can thoroughly memorize the data distribution of this large-scale dataset, leading to overfitting. Increasing numbers of model parameters may further exacerbate this problem. Therefore, the degradation of the generalization performance of facial landmark detectors in unconstrained environments may not be addressed by simply increasing human annotations and model parameters.

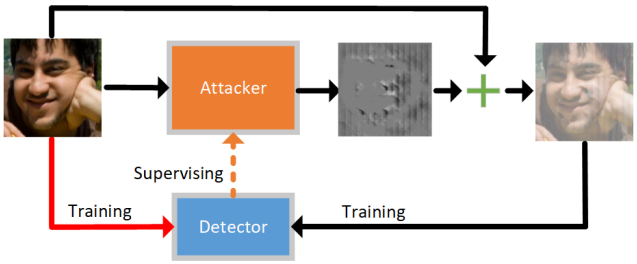

Figure 1. Insight of the proposed SAAT. The attacker generates adversarial perturbations to fool the detector, and the detector then learns to defend against attacks.

In recent years, researchers have proposed various data augmentation methods [8, 10, 12, 30, 49] to achieve robust facial landmark detection. Wingloss [12] balances the data distribution with handcrafted transformations (flipping, rotation, scaling *etc*.). However, these rigid transformations are not adequate for those detectors facing various attacks from the real world (illumination, make-ups, skin color,

*Corresponding author.

etc.). More recently, unsupervised learning is introduced to improve the generalization of models by automatically annotating large-scale unlabeled data [10, 48, 49] while it requires prior knowledge provided by a pre-trained detector on the existing labeled data. Because the annotated new samples may heavily overlap with 'easy' samples in pre-training data, unsupervised methods [10, 48, 49] can hardly maintain accurate detection on 'hard' samples. To diversify face data, a few methods [8, 30] introduce face augmentation (*e.g*., style transfer or face generation). Style aggregation [8] augments faces in the aggregated style, which is a test-time augmentation. This method increases computational cost at the testing phase, and its style transfer is limited to three handcrafted styles. AVS [30] augments "new" faces by leveraging generative adversarial networks (GAN). However, it may also generate unrealistic faces owing to the high uncertainty of the GAN models and the complexity of face generation. In addition, the performance gains of unsupervised learning methods [48, 49] and face augmentation methods [8, 30] are dependent on their pre-training models, resulting in their generalization performance being limited by pre-training. Unfortunately, Goodfellow *et al*. [14, 20, 29] proved that deep neural networks trained on even large datasets are vulnerable to human-imperceptive perturbations.

We propose a sample-adaptive adversarial training (SAAT) approach to address the challenges mentioned above, which exploits adversarial attacks to enhance detectors, as shown in Figure 1. In this framework, an attacker generates adversarial perturbations instead of "new" faces to fool a facial landmark detector, and the detector learns to defend against these perturbations. Generally, most existing attacks [14, 20, 29] generate category-specific adversarial perturbations that are usually human-imperceptible. They always have full access to a pre-trained model and use various gradient-based methods to generate adversarial examples. Unlike these attacks, we design a sample-adaptive black-box attacker, which does not require any pre-trained model, but rather crafts adversarial perturbations based on the real-time performance feedback of the detector running on current adversarial examples. In addition, we allow the perturbations to be visible to diversify further adversarial examples, which helps the detector learn to defend against visible attacks from the real world (illumination, make-up, skin color, *etc*.). Then the adversarial perturbations must avoid the generation of false cases (*e.g*. faces with three eyes or twisted mouths) and structural inconsistency of face shapes between the adversarial and corresponding original faces. To achieve this, the attacker induces adversarial perturbations to adapt to different faces. Specifically, the attacker implicitly learns the face structure of the adversarial examples by leveraging a structure-guided conditional adversarial architecture. Moreover, a semantic reconstruction loss is employed to explicitly constrain the semantic consistency between adversarial examples and the corresponding original faces. Nevertheless, false samples may still be generated during end-to-end training, especially before the convergence of the model, owing to these visible perturbations. To avoid irreversible degradation caused by false samples, we introduce a sample-adaptive weight, which automatically adjusts the contribution of different adversarial examples to the detector in each training step by measuring the structural similarity of the adversarial examples and the corresponding original faces. Our main contributions are summarized as follows:

- The proposed approach performs robust facial landmark detection as a defense against attacks from the real world. It can improve the robustness of facial landmark detectors without increasing model parameters and human annotations.

- Based on facial semantic information, the proposed sample-adaptive black-box attacker induces visible perturbations to adapt to different faces by interacting with the detector. It injects these perturbations into the training data, complementing existing data augmentations to reduce the risk of overfitting.

- We introduce a sample-adaptive weight to avoid the detector's performance degradation caused by false samples. This weight allows the attacker and detector to be interactively optimized in end-to-end training without any pre-training.

2. Related work

Facial landmark detection, *a.k.a.* face alignment, aims to localize the key points in a given face image. There are a number of classic and effective approaches for this task [7, 10, 15, 36, 40, 41, 51]. In [5, 37, 41, 51], researchers addressed face alignment as a cascaded regression process, which refines the initial shape to the final shape in a coarse-to-fine manner. After the application of CNNs in this field, works such as [11, 16, 25, 36, 26, 44] achieved competitive performance, they learn discriminative features from pixels. With the large pose issues taken into consideration, 3D face fitting has been considered [1, 15, 18, 21], which aims to fit a 3D morphable model (3DMM) to a 2D image. Wing loss [12] designs a piece-wise loss function, which amplifies the impact of the errors from a certain interval by switching from L1 loss to a modified logarithm function. To develop the more powerful detector, the heatmap regression methods [6, 3, 10, 24, 28, 40] are presented. These methods achieved breakthroughs of facial landmark detection by stacking model parameters. Look at boundary (LAB) [40] uses stacked hourglass networks to impose

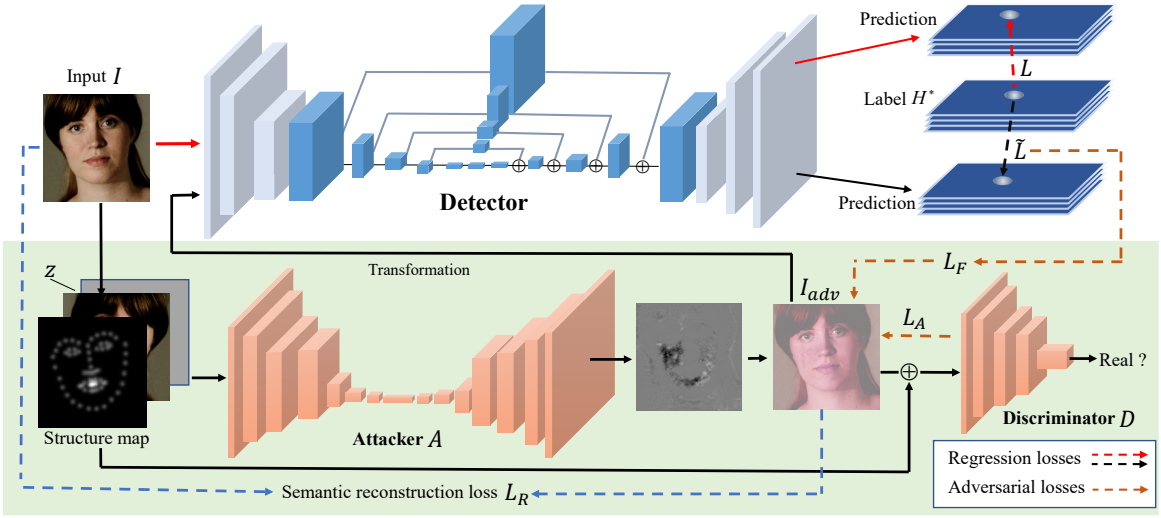

Figure 2. Illustration of the proposed method. The structure-guided attacker A generates adversarial perturbations that are injected into the face image I, which synthesizes the adversarial example I_{adv}. The detector then learns to handle both I and I_{adv}. We optimize A by jointly minimizing L_F, L_A, and L_R. L_F aims to increase the prediction error of the detector on adversarial example I_{adv}. L_A aims to implicitly learn an awareness of the face structures contained in the training data. The semantic reconstruction loss L_R explicitly maintains semantic consistency between the adversarial example and the original face. L and $\tilde{L}$ denote the regression losses of the detector running on the original and adversarial example, respectively. z represents the Gaussian noise, and $\oplus$ denotes the concatenation operation.

the global geometric constraint over all landmarks by introducing boundary information. However, LAB is computationally expensive due to the boundary heatmap prediction. SA [24] proposes semantic alignment, which reduces the 'semantic ambiguity' intrinsically and stacks additional sub-networks and multi-stage hourglass networks to impose the shape constraint. Although the stacked additional sub-networks improves the fitting ability, it also raises the risk of overfitting. To augment the annotations, methods such as [10, 48, 49] apply self-supervised learning to annotated large unlabeled video data. SBR [10] employs the Lucas-Kanade operation that is sensitive to light change and occlusion. STRRN and STKI [49, 48] are heavily dependent on their pre-training detector, resulting in that their generalization performances are also limited by pre-training.

Defense against adversarial attacks: Szegedy [34] first demonstrated that the perturbed images could fool deep neural networks into misclassification, and the robustness of deep neural networks against the adversarial examples could be improved by adversarial training. Subsequently, Goodfellow and Bengio *et al*. [14, 29, 20] shown that deep neural networks even trained on the large data are vulnerable to human-imperceptive adversarial perturbations. To improve the robustness of the neural networks, Virtual Adversarial Training approach *et al*. [27] applied perturbations to the word embedding to smooth the output distributions of the neural networks. A related work [47] also proposed the 'stability training' method to improve the robustness of neural networks against small distortions to input images.

3. Method

Our framework consists of two competitive-cooperative parts, a replaceable facial landmark detector and a conditional generator as the attacker A. The attacker produces adversarial examples by injecting perturbations into inputs to fool the detector, and then the detector learns to handle these adversarial examples, see Figure 2.

3.1. Replaceable detector

We use a non-stacked hourglass network (HG*1) as our detector in this section. Given an image I, let ϕ represent the prediction process in which the detector maps I to a landmark heatmap set H, such that $H = \phi(I)$. We optimize the detector by minimizing the following $L2$ loss:

$$L = \frac{1}{N}\sum_{n=1}^{N} \|H_n - H_n^*\|^2, \tag{1}$$

where N denotes the number of landmark heatmaps and H^* is the ground-truth heatmap set.

3.2. Sample-adaptive adversarial attacks

We allow the input face I to be modified by generating visible adversarial perturbations. This will produce an adversarial example I_{adv}, which we consider as the hard sample. We design a structure-guided conditional generator as the attacker A to generate adversarial perturbations to maintain a reasonable face structure. Guided by the structure

map S of face I, attacker A can quickly explore the areas expected to be attacked, which perturbs the input I to synthesize an adversarial example I_{adv}:

$$I_{adv} = I + A\left(I, S, z\right) \tag{2}$$

where z represents the Gaussian noise, and S is obtained by mapping all landmark heatmaps to a single map. This sample aims to reflect the weaknesses of the detector. Based on the performance feedback of the detector running on the current adversarial example, A is optimized by minimizing the following loss:

$$L_F = \frac{1}{\frac{1}{N}\sum_{n=1}^{N}\left\|\phi\left(I_{adv}\right)_n - H_n^*\right\|^2 + \lambda}, \tag{3}$$

where $\lambda = 0.1$ denotes the relaxation factor. L_F encourages the attacker to generate various adversarial perturbations that adapt to different faces according to the feedback of the current perturbations. Attacker A must produce a harder sample when the detector is able to handle the adversarial example by learning a defense against attacks. This competitive game prevents the detector from overfitting but may also lead to over-attack, generating false samples. λ is introduced to constrain the change in the loss value within a reasonable interval, avoiding over-attack.

To avoid generating false samples (three eyes, twisted mouth, *etc*.), we force the synthesized face to be real. To achieve this, we use adversarial supervision to update the attacker by minimizing the following loss:

$$L_A = \mathbb{E}_{I,S,z}\left[\log\left(1 - D\left(A\left(I, S, z\right) + I, S\right)\right)\right]. \tag{4}$$

The discriminator D supervises the attacker to learn an awareness of the structural consistency between the adversarial example and the structure map. To this end, we maximize the following loss to optimize the discriminator:

$$\begin{aligned} L_D = &\mathbb{E}_{I,S}\left[\log D\left(I, S\right)\right] + \\ &\mathbb{E}_{I,S,z}\left[\log\left(1 - D\left(A\left(I, S, z\right) + I, S\right)\right)\right]. \end{aligned} \tag{5}$$

Adversarial learning implicitly imposes the face structure constraint into the adversarial example.

To ensure that the annotations of the original data are available for adversarial examples that are used to train the detector, the attacker must retain the facial semantic information of the adversarial example to be consistent with the input face. To achieve this, we segment the neighborhoods of facial landmarks as semantic regions and the remainder as non-semantic regions. A semantic mask M performs this segmentation. Specifically, the proposed method maps all landmark heatmaps to the semantic mask M and reset the value of M by setting the values of the semantic regions to 1 and the non-semantic regions to 0.1. Thus, the semantic reconstruction loss is defined as follows:

$$L_R = \left\|M * I - M * I_{adv}\right\|_1. \tag{6}$$

The semantic reconstruction loss maintains semantic consistency between the adversarial example and the original input. This loss allows the ground-truth heatmaps to be available for the corresponding adversarial examples. Therefore, adversarial examples can be used to optimize the detectors.

Finally, we combine all losses of the attacker as the total adversarial loss:

$$L_T = L_F + L_A + L_R. \tag{7}$$

This total loss encourages the attacker to suitably perturb training data to produce adversarial examples that can improve the robustness of the detector against attacks from the real world. Note that the parameters of the detector are fixed when the attacker is updated using L_T.

3.3. Sample-adaptive training strategy

We expect the adversarial examples to be sufficiently credible to train the detector. Nevertheless, because the perturbations change the context of the original faces, the attacker may produce false samples before the convergence to degrade the detector. Most of the existing adversarial attacks address this problem by limiting the values of adversarial perturbations to a small range. However, this restricts the data diversity of adversarial examples. Because we also expect the adversarial examples to be diversified, our attacker generates visible perturbations. It is necessary to balance the benefits and risks of visual perturbations. Therefore, we design a sample-adaptive loss to efficiently exploit all credible adversarial examples whose semantic regions are consistent with the corresponding original faces. We express the adaptive sample loss as:

$$\tilde{L} = \beta * \frac{1}{N}\sum_{n=1}^{N}\left\|\phi(I_{adv})_n - H_n^*\right\|^2, \tag{8}$$

where N denotes the number of landmarks, and the weight β aims to balance the risks and benefits of the visible perturbations. Based on the structural similarity [38] between the adversarial example and the corresponding original face in semantic regions, this weight evaluates the confidence level of the adversarial example to adjust $\tilde{L}$. We define β as:

$$\beta = \max\left(0, \ln\left(SSIM(M * I, M * I_{adv}) + 0.5\right)\right), \tag{9}$$

where $SSIM(\cdot)$ measures the structural similarity [38], M represents the weight mask of the semantic regions.

In a single training step, we combine the loss L and the loss $\tilde{L}$ to optimize the detector, and the parameters of the attacker are fixed when the detector is updated. Moreover, we also augment the adversarial example with handcrafted transformation (rotation, scaling, flipping, *etc*.)

4. Experiments

4.1. Datasets

Our method is evaluated in three settings, including intra-dataset evaluation, cross-dataset evaluation and self-evaluation.

Intra-dataset evaluation setting follows the settings of most methods [10, 23, 36, 40] on 300W [31], 300VW [33], and WFLW [40]:

300W [31] contains three training subsets, which are the training set of LFPW ($2,000$ images), the training set of HELEN (811 images) and AFW (337 images). Following the widely used evaluation setting, the test sets consist of the Common set (LFPW and HELEN test sets, 554 images), the Challenging set (IBUG, 135 images), and the Full set (the union of the two former, 689 images).

300VW [33] contains 114 videos. Following most methods [22, 35, 40], 50 training videos and training sets of 300W are combined during the training phase. The remainder 64 videos are used for testing, which are divided into three subsets Cate-1, Cate-2 and Cate-3 (challenging set).

WFLW [40] is a new facial dataset based on WIDER Face [43], which is proposed by LAB [40]. It contains $10,000$ images ($7,500$ for training and $2,500$ for testing) with 98 landmarks.

Cross-dataset evaluation setting aims to demonstrate the generalization performance of the proposed method. We only use the train sets of 300W [31] to train our model in this setting without using any samples from other datasets. Then the detector is tested on the following datasets:

COFW68 [13] has 507 test images collected from the Internet, which is produced based on The Caltech Occluded Face in the Wild (COFW) dataset [4]. Since COFW are annotated with 29 landmarks, it can not evaluate the detector trained on 300W with 68 landmarks. In this setting, we conduct evaluation experiments on COFW68 re-annotated with 68 landmarks by [13].

Masked 300W is a 300W based dataset focusing on masked faces. Although COFW68 contains various occluded faces, it lacks severe occlusions. We synthesize masked face images to generated the Masked 300W dataset by following Simulated Masked Face Recognition Dataset (SMFRD) [39]. More details are illustrated in our supplementary material.

Self-evaluation setting is used in the efficiency evaluation of our SAAT. We report the role of multiple hyper-parameters and analyze the impact of SAAT on different baseline and data sizes. We train different sizes of hourglass networks on HELEN, 300W [31] and a large scale 300W-LP datasets [3] ($61,225$), respectively, where 300W-LP is usually used to pre-train detectors.

4.2. Implementation details

Model details. We apply three baselines to evaluate our SAAT, respectively. The cascade regression model MDM[1] [36] are used as baseline-1. The non-stacked HG model is baseline-2 (HG*1), which only equips a non-stacked hourglass network with a single hourglass module [28]. It starts with a $7 * 7$ convolution layer with a stride of 2, followed by a round of residual module and max-pooling to bring the resolution down from 256 to 64. The two subsequent residual modules precede the hourglass module, followed by two deconvolution layers to output a $68 * 256 * 256$ feature. The final two convolution layers with $1 * 1$ kernel are equipped to generate the heatmaps. To further evaluate the proposed method, we also apply two-stage stacked HG (HG*2) as baseline-3. The structure-guided conditional GAN follows pix2pix [2] [17] to be completed.

Implementation details. We use the face bounding boxes released by organizers to crop the face image of the samples in 300W. MTCNN [45] is applied to detect the face bounding boxes for 300VW. All images are cropped into a size of $256 * 256$. Following the existing methods [12, 36, 40], we augment training data by handcrafted transformation (rotation, flipping, scaling, *etc*.). Our model is trained with about $200,000+$ steps in an end-to-end manner on a single GPU with the NVIDIA GTX 1080 Ti. The following hyper-parameters are set: a initial learning rate of 0.0002 for the detector, a initial learning rate of 0.0001 for the attacker and discriminator, a decay factor of 0.97 and a batch size of 8.

Evaluation metric. We use the point-to-point Euclidean error normalized by the inter-ocular distance [31, 36] to produce the Normalized Mean Error (NME). We further leverage the Cumulative Error Distribution (CED) curves of NMEs to quantitatively evaluate the performance. Although the averaged NME and CED curves are widely used for evaluation, they are insensitive to a few false cases. Therefore, we also provide our evaluation in the form of the visualization of all NMEs.

4.3. Intra-dataset evaluation

For fair comparisons to previous methods, we re-produce the results with their released codes or present the reported results from their original papers [2, 9, 10, 23, 25, 40, 41, 50]. For the CED curves, we obtain the results with publicly released codes [10, 28, 35, 36, 23, 40].

Evaluation on 300W. Table 1 shows that the proposed SAAT significantly improves all baselines on three sub-sets and even surpasses the newest methods by a large margin on the challenging set. Although Chandran *et al.*[6] is competitive on the common set, it can hardly handle the chal-

[1] https://github.com/trigeorgis/mdm

[2] https://github.com/yenchenlin/pix2pix-tensorflow

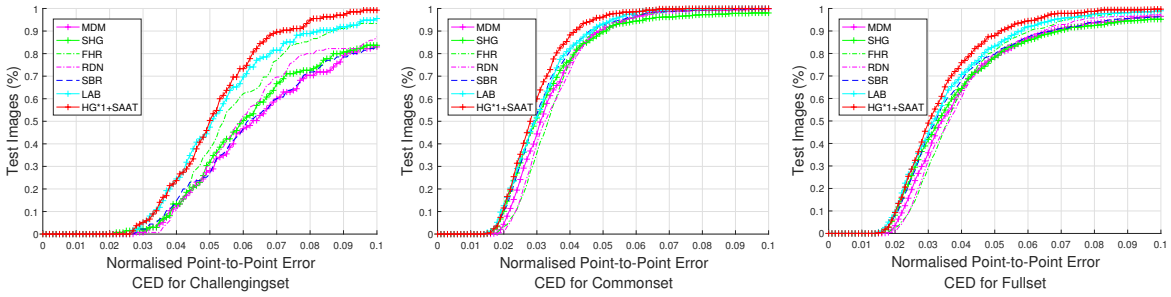

Figure 3. The CED curves of our proposed method compared to state-of-the-art methods on three subsets of 300W [31].

Methods	Challenging	Common	Full
CFSS [51] (2015)	9.98	4.73	5.76
SHG [28] (2016)	7.52	3.17	4.01
TSR [25] (2017)	7.56	4.36	4.99
RDN [23] (2018)	7.04	3.31	4.23
SBR [10] (2018)	7.58	3.28	4.10
SHG+Wing [12] (2018)	5.64	3.05	3.56
FHR [35] (2018)	6.28	3.02	3.66
LAB [40](2018)	5.19	2.98	3.49
ODN [50] (2019)	6.67	3.56	4.17
AVS+SAN [30] (2019)	6.49	3.21	3.86
Chandran *et al*. [6](2020)	7.04	2.83	4.23
LUVLi [19] (2020)	5.16	**2.76**	**3.23**
3FabRec [2] (2020)	5.74	3.36	3.82
SRT [9] (2020)	5.61	**2.80**	3.39
MDM [36] (baseline-1)	8.87	3.74	4.78
HG*1 (baseline-2)	7.13	3.50	4.21
HG*2 (baseline-3)	6.43	3.11	3.76
MDM+SAAT	6.58	3.35	3.98
HG*1+SAAT	**5.10**	2.96	3.38
HG*2+SAAT	**5.03**	2.87	**3.29**

Table 1. Comparison with the state-of-the-arts on 300W. HG*1 (baseline-2) represents a single hourglass network, and HG*2 (baseline-3) denotes a two-stage stacked hourglass network. Top-2 results are highlighted in bold font.

Methods	Cate-1	Cate-2	Cate-3
Trained on 300W + 300VW			
TCDCN [46]	7.66	6.77	14.98
MDM [36]	5.46	4.59	7.42
iCCR [32]	6.71	4.00	12.75
TSTN [22]	5.36	4.51	12.84
FHR+STA [35]	4.42	4.18	5.98
HGs+SA+GHCU [24]	3.85	3.46	7.51
HG*1 (baseline-2)	4.01	4.12	6.67
HG*1+SAAT (ours)	**3.46**	**3.41**	**5.23**
Trained on 300W + Unlabeled data			
SBR [10]	7.41	6.18	11.04
STRRN [49]	5.03	4.74	6.63
STKI [48]	4.42	4.57	6.11
Trained on 300W			
MDM [36]	6.76	5.69	8.74
SHG [28]	4.68	4.41	8.13
HG*1 (baseline-2)	4.71	4.35	8.37
MDM+SAAT (ours)	5.96	4.95	8.74
HG*1+SAAT (ours)	4.03	3.87	6.12

Table 2. NME comparison of the proposed method and state-of-the-art methods on the 300VW.

lenging set. This is because this model relies on the stacked hourglass networks to crop the different face regions. However, many hard samples with the large pose in the challenge set can hardly be cropped well. The newest methods SRT [9] also achieves outstanding performance on the common set. However, it does not perform well on the challenging set. Because SRT automatically annotates large-scale unlabeled video data by exploiting Lucas Kanade tracking, which can hardly handle occlusion and illumination, leading to data imbalance. Although LUVLi [19] is successful on 300W, it is computationally expensive as it is equipped with 8 U-nets. Our proposed method achieves the best performance on the challenging set, which shows that the proposed SAAT can significantly improve the robustness of the detectors on hard samples.

Figure 3 shows the CED curves of HG*1+SAAT compared with state-of-the-art methods whose mean error values are presented in Table 1. It is seen that the HG*1+SAAT significantly surpasses other methods.

Evaluation on 300VW. Comparison results are report in Table 2. We can see that the proposed method significantly outperforms other face alignment methods on all sub-sets. Even on the challenging set Cate-3 with large poses, diverse expressions and severe occlusions, our method still achieves outstanding performance.

Evaluation on WFLW. To further evaluate the robustness of our method, we report mean error, failure rate and AUC (a threshold value of 0.1) on the Testset of WFLW in Table 3. Compared with the top-1 LUVLi [19], although our detector uses fewer parameters, it still achieves a compelling performance.

Methods	LAB (HG*4) [40]	AS (Res-18) [30]	3FabRec (more data) [2]	LUVLi (U-net*8) [19]	HG*1+SAAT (ours)	HG*2+SAAT (ours)
Mean Error (%)	5.27	5.25	5.62	4.37	5.37	5.11
Failure Rate (%)	7.56	7.44	8.28	3.12	7.31	5.63
AUC	0.5323	0.5034	0.4840	0.5770	0.5201	0.5633

Table 3. Evaluation on Testset of WFLW (98 landmarks).

Method	NME	Failure Rate (%)
RPP [42]	7.52	9.20
CFAN [44]	8.38	7.14
TCDCN [46]	8.05	6.31
MDM [36]	6.32	4.31
FAN [3]	5.72	2.74
ODN [50]	5.87	2.84
LAB [40]	4.62	2.17
HG*1+SAAT (ours)	**4.61**	**1.58**

Table 4. Comparison of NME and the failure rate (threshold at 0.1) on COFW68 dataset.

4.4. Cross-dataset evaluation

In this evaluation, we only use 300W to train our model.

Evaluation on COFW68. We conduct cross-dataset experiments on COFW68 to evaluate the robustness. Table 4 shows the comparison results on COFW68. The proposed method outperforms all state-of-the-art methods by a large margin. This also proves that adversarial examples are helpful to promote localization performance.

Evaluation on Masked 300W. Table 5 shows results of the proposed SAAT compared with existing open-sourced methods on Masked 300W. For this evaluation, we re-train these methods without using masked faces from Masked 300W dataset, but using the handcrafted occlusion patch to randomly augment occluded data. Our SAAT improves baselines significantly without increasing computational cost at the testing phase. This shows that our method can efficiently complement the handcrafted data augmentation.

Methods	Challenging	Common	Full
CFSS [51]	19.98	11.73	13.35
SBR [10]	15.28	9.72	10.65
SHG [28]	13.52	8.17	9.22
FHR [35]	12.38	7.82	8.71
FAN [3]	10.81	7.36	8.02
MDM [36](baseline-1)	11.67	7.66	8.44
HG*1(baseline-2)	16.18	9.19	10.56
HG*2(baseline-3)	14.46	9.03	10.94
MDM+SAAT	10.78	6.93	7.68
HG*1+SAAT	12.58	6.37	7.58
HG*2+SAAT	11.36	5.42	6.58

Table 5. NME comparison on Masked 300W. Note that Masked 300W is only used for testing, not training.

Methods	Challenging	Common	Full
HG*1+SAAT ($\beta = 1$)	6.36	3.47	4.04
HG*1+SAAT ($\beta = 0.5$)	5.78	3.19	3.70
HG*1+SAAT ($\beta = 0.1$)	6.91	3.32	4.02
HG*1+SAAT ($\beta = 0$)	7.93	3.60	4.45
HG*1+SAAT ($\lambda = 1$)	6.13	3.66	4.14
HG*1+SAAT ($\lambda = 0.5$)	5.17	3.01	3.43
HG*1+SAAT ($\lambda = 0$)	7.06	3.42	4.13
HG*1+SAAT ($\beta = 0, \lambda = 0$)	8.72	4.04	4.95
HG*1+SAAT (Adaptive β, $\lambda = 0.1$)	5.10	2.96	3.38

Table 6. Ablation experiment: NME comparison on 300W dataset.

4.5. Self Evaluations

Ablation study. We show averaged errors of various training strategies with different values of β and λ in Table 6, respectively. We can observe two conclusions from these results: 1) Adaptive β achieves higher performance since it efficiently filters the false samples, and 2) λ can avoid the increase of false samples caused by the over-attack. We argue that these two parameters cooperatively work to suppress false samples.

Table 7 shows the ablation experiments about the structure map S, the reconstruction loss L_R and the discriminator D. A single hourglass network is applied as the baseline. More ablation studies about the quality of adversarial examples are shown in our supplementary material.

300W validation set	Challenging	Common	Full
SAAT (w/o S)	5.36	3.17	3.78
SAAT (w/o L_R)	6.13	3.32	3.99
SAAT (w/o D)	5.68	3.01	3.53
SAAT (w/o D and L_R)	7.51	4.09	4.76
SAAT	5.10	2.96	3.38

Table 7. Ablation experiments of the proposed SAAT.

Evaluation on hard samples. Due to the tiny scale of hard samples in 300W validation sets, the averaged statistical results are unable to clearly reflect the performance of the detector running on hard samples. We further investigate the performance of the proposed SAAT in handling hard samples. Figure 4 shows the errors of the baseline and baseline+SAAT on all samples of 300W validation sets. Although the baseline fits most samples, a few hard samples cannot be fitted by the baseline. The proposed method addresses this problem. Figure 5 shows the qualitative results. These results indicate that our SAAT can efficiently improve the baseline's robustness.

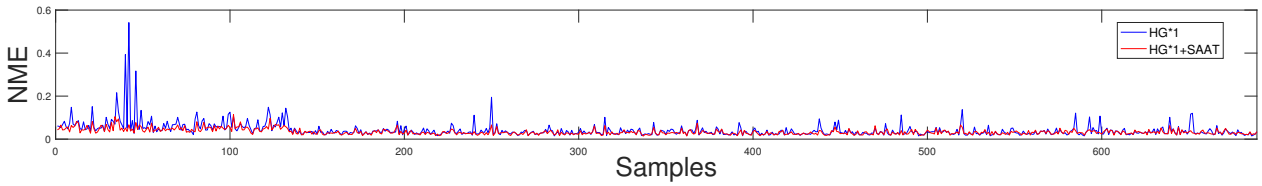

Figure 4. Quantitative results of the proposed SAAT compared with the baseline on 300W.

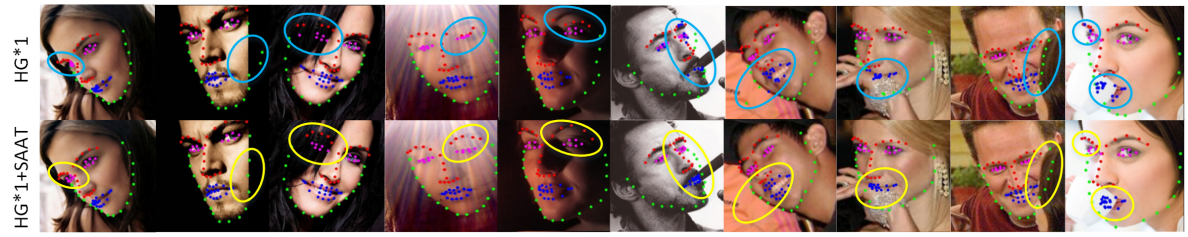

Figure 5. Qualitative results of the proposed SAAT compared with the baseline on hard samples.

Methods	LAB [40]	SBR [10]	SHG+GHCU [24]	FAN [3]	HG*1+SAAT
Stacks	4	3	4	4	1
FPS	9.63	13.82	8.79	13.2	23.3
Sub-nets	✓	×	✓	×	×

Table 8. Comparison of the number of stacked hourglass modules, inference speed and sub-networks at the testing phase. FPS denotes the number of images processed per second and Sub-nets denotes extra modules in addition to stacked models.

Model size and computational efficiency. We further compare our HG*1+SAAT (baseline-2) with the stacked models on the number of stacked hourglass modules, the inference speed, and extra sub-nets used to the testing phase. All experiments are conducted on a single NVIDIA GTX 1080 Ti and reported in Table 8. Combining inter-dataset evaluation (Table 1, 2, 3) and cross-dataset evaluation (Table 4, 5), these results shown that the proposed SAAT can improve the performance of detectors without increasing model parameters.

Performance across different network and data sizes. In Table 9, we follow FAN [3] to report the performance fluctuation of the proposed SAAT with different numbers of network parameters. Specifically, we train the proposed SAAT using three different data sizes separately, including: HELEN (811 images), 300W ($3,148$ images), and 300W+300W-LP [3] ($61,225$ images), and test on COFW68 [13], respectively. The number of stacked hourglass modules is varied from 2 to 1. These results are collected after $250,000$ training steps, where the observation interval was 3000 training steps. The range of NME fluctuation after convergence can measure the model's generalization. The greater fluctuation is usually caused by severer overfitting. These results show that the proposed SAAT is better for the limited training data and computational resources.

Trained on	HG*1	HG*2	HG*1+SAAT	HG*2+SAAT
HELEN	5.56-6.82	5.52-21.79	5.26-6.77	5.12-9.31
300W	5.13-7.74	4.86-8.33	4.81-6.23	4.71-7.59
300W+300W-LP	4.76-6.01	4.53-6.09	4.69-6.17	4.47-5.73

Table 9. NME fluctuation of different model sizes on the COFW.

5. Conclusion

In this paper, we have proposed a sample-adaptive adversarial training (SAAT) approach, which significantly improves the robustness of facial landmark detection. Extensive experimental results confirmed that the proposed SAAT achieves competitive performance compared with state-of-the-art methods. The advantages of the proposed method are as follows: (1) SAAT exploits diverse adversarial examples to reduces the overfitting risk; (2) With the defense against visible adversarial attacks, the detector can achieve robust detection; (3) the proposed sample-adaptive weight can suppress false samples, which avoids the performance degeneration. In the future, we will consider promoting SAAT to related tasks, such as human pose estimation.

6. Acknowledgements

This work is supported in part by Shanghai science and technology committee under grant No.21511100600. We appreciate the High Performance Computing Center of Shanghai University, and Shanghai Engineering Research Center of Intelligent Computing System (No.19DZ2252600) for providing the computing resources and technical support.

References

[1] Volker Blanz and Thomas Vetter. Face recognition based on fitting a 3d morphable model. *TPAMI*, 25(9):1063–1074, 2003. 2

[2] Bjorn Browatzki and Christian Wallraven. 3fabrec: Fast few-shot face alignment by reconstruction. In *CVPR*, June 2020. 5, 6, 7

[3] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *ICCV*, pages 1021–1030, 2017. 1, 2, 5, 7, 8

[4] Xavier P. Burgos-Artizzu, Pietro Perona, and Piotr Dollar. Robust face landmark estimation under occlusion. In *ICCV*, December 2013. 1, 5

[5] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. Face alignment by explicit shape regression. *IJCV*, 107(2):177–190, 2014. 2

[6] Prashanth Chandran, Derek Bradley, Markus Gross, and Thabo Beeler. Attention-driven cropping for very high resolution facial landmark detection. In *CVPR*, pages 5861–5870, 2020. 2, 5, 6

[7] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. *TPAMI*, 23(6):681–685, 2001. 2

[8] Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Style aggregated network for facial landmark detection. In *CVPR*, pages 379–388, 2018. 1, 2

[9] X. Dong, Y. Yang, S. Wei, X. Weng, Y. Sheikh, and S. Yu. Supervision by registration and triangulation for landmark detection. *TPAMI*, pages 1–1, 2020. 5, 6

[10] Xuanyi Dong, Shoou-I Yu, Xinshuo Weng, Shih-En Wei, Yi Yang, and Yaser Sheikh. Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors. In *CVPR*, pages 360–368, 2018. 1, 2, 3, 5, 6, 7, 8

[11] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. *arXiv:1803.07835*, 2018. 2

[12] Zhen-Hua Feng, Josef Kittler, Muhammad Awais, Patrik Huber, and Xiao-Jun Wu. Wing loss for robust facial landmark localisation with convolutional neural networks. In *CVPR*, pages 2235–2245, 2018. 1, 2, 5, 6

[13] Golnaz Ghiasi and Charless C. Fowlkes. Occlusion coherence: Detecting and localizing occluded faces. *CoRR*, abs/1506.08347, 2015. 5, 8

[14] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 2, 3

[15] Shi HL et al. Face alignment across large poses: A 3d solution. In *CVPR*, pages 146–155, 2016. 2

[16] Zhibin Hong, Xue Mei, Danil Prokhorov, and Dacheng Tao. Tracking via robust multi-task multi-view joint sparse representation. In *ICCV*, pages 649–656, 2013. 2

[17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 1125–1134, 2017. 5

[18] Amin Jourabloo and Xiaoming Liu. Pose-invariant face alignment via cnn-based dense 3d model fitting. *IJCV*, 124(2):187–203, 2017. 2

[19] Abhinav Kumar, Tim K Marks, Wenxuan Mou, Ye Wang, Michael Jones, Anoop Cherian, Toshiaki Koike-Akino, Xiaoming Liu, and Chen Feng. Luvli face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood. In *CVPR*, pages 8236–8246, 2020. 1, 6, 7

[20] Alexey Kurakin, Ian Goodfellow, Samy Bengio, et al. Adversarial examples in the physical world. 2, 3

[21] Feng Liu, Dan Zeng, Qijun Zhao, and Xiaoming Liu. Joint face alignment and 3d face reconstruction. In *ECCV*, pages 545–560. Springer, 2016. 2

[22] Hao Liu, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Two-stream transformer networks for video-based face alignment. *TPAMI*, 40(11):2546–2554, 2018. 5, 6

[23] Hao Liu, Jiwen Lu, Minghao Guo, Suping Wu, and Jie Zhou. Learning reasoning-decision networks for robust face alignment. *TPAMI*, 2018. 5, 6

[24] Zhiwei Liu, Xiangyu Zhu, Guosheng Hu, Haiyun Guo, Ming Tang, Zhen Lei, Neil M Robertson, and Jinqiao Wang. Semantic alignment: Finding semantically consistent ground-truth for facial landmark detection. In *CVPR*, pages 3467–3476, 2019. 1, 2, 3, 6, 8

[25] Jiangjing Lv, Xiaohu Shao, Junliang Xing, Cheng Cheng, and Xi Zhou. A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In *CVPR*, pages 3317–3326, 2017. 1, 2, 5, 6

[26] Daniel Merget, Matthias Rock, and Gerhard Rigoll. Robust facial landmark detection via a fully-convolutional local-global context network. In *CVPR*, pages 781–790, 2018. 2

[27] Takeru Miyato, Andrew M Dai, and Ian Goodfellow. Adversarial training methods for semi-supervised text classification. *ICLR*, 2017. 3

[28] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, pages 483–499. Springer, 2016. 1, 2, 5, 6, 7

[29] 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, 3

[30] Shengju Qian, Keqiang Sun, Wayne Wu, Chen Qian, and Jiaya Jia. Aggregation via separation: Boosting facial landmark detector with semi-supervised style translation. In *CVPR*, pages 10153–10163, 2019. 1, 2, 6, 7

[31] 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, 6

[32] Enrique Sánchez-Lozano, Brais Martínez, Georgios Tzimiropoulos, and Michel F. Valstar. Cascaded continuous regression for real-time incremental face tracking. In *ECCV*, pages 645–661, 2016. 6

[33] Jie Shen, Stefanos Zafeiriou, Grigoris G. Chrysos, Jean Kossaifi, Georgios Tzimiropoulos, and Maja Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *ICCVW*, pages 1003–1011, 2015. 1, 5

[34] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 3

[35] Ying Tai, Yicong Liang, Xiaoming Liu, Lei Duan, Jilin Li, Chengjie Wang, Feiyue Huang, and Yu Chen. Towards highly accurate and stable face alignment for high-resolution videos. In *AAAI*, volume 33, pages 8893–8900, 2019. 1, 5, 6, 7

[36] George Trigeorgis, Patrick Snape, Mihalis A Nicolaou, Epameinondas Antonakos, and Stefanos Zafeiriou. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *CVPR*, pages 4177–4187, 2016. 1, 2, 5, 6, 7

[37] Georgios Tzimiropoulos. Project-out cascaded regression with an application to face alignment. In *CVPR*, pages 3659–3667, 2015. 2

[38] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 13(4):600–612, 2004. 4

[39] 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. 5

[40] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *CVPR*, pages 2129–2138, 2018. 1, 2, 5, 6, 7, 8

[41] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, pages 532–539, 2013. 1, 2, 5

[42] Heng Yang, Xuming He, Xuhui Jia, and Ioannis Patras. Robust face alignment under occlusion via regional predictive power estimation. *TIP*, 24(8):2393–2403, 2015. 1, 7

[43] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *CVPR*, pages 5525–5533, 2016. 5

[44] Jie Zhang, Shiguang Shan, Meina Kan, and Xilin Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *ECCV*, pages 1–16, 2014. 2, 7

[45] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. 5

[46] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Learning deep representation for face alignment with auxiliary attributes. *TPAMI*, 38(5):918–930, 2016. 1, 6, 7

[47] Stephan Zheng, Yang Song, Thomas Leung, and Ian Goodfellow. Improving the robustness of deep neural networks via stability training. In *CVPR*, pages 4480–4488, 2016. 3

[48] Congcong Zhu, Xiaoqiang Li, Jide Li, Guangtai Ding, and Weiqin Tong. Spatial-temporal knowledge integration: Robust self-supervised facial landmark tracking. In *ACM International Conference on Multimedia*, page 4135–4143. Association for Computing Machinery, 2020. 2, 3, 6

[49] Congcong Zhu, Hao Liu, Zhenhua Yu, and Xuehong Sun. Towards omni-supervised face alignment for large scale unlabeled videos. In *AAAI*, pages 13090–13097, 2020. 1, 2, 3, 6

[50] Meilu Zhu, Daming Shi, Mingjie Zheng, and Muhammad Sadiq. Robust facial landmark detection via occlusion-adaptive deep networks. In *CVPR*, pages 3486–3496, 2019. 5, 6, 7

[51] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang. Face alignment by coarse-to-fine shape searching. In *CVPR*, pages 4998–5006, 2015. 1, 2, 6, 7