ADNet: Leveraging Error-Bias Towards Normal Direction in Face Alignment

Yangyu Huang, Hao Yang*, Chong Li, Jongyoo Kim, Fangyun Wei

Microsoft Research Asia

{yanghuan,haya,chol,jongk,fawe}@microsoft.com

Abstract

The recent progress of CNN has dramatically improved face alignment performance. However, few works have paid attention to the error-bias with respect to error distribution of facial landmarks. In this paper, we investigate the error-bias issue in face alignment, where the distributions of landmark errors tend to spread along the tangent line to landmark curves. This error-bias is not trivial since it is closely connected to the ambiguous landmark labeling task. Inspired by this observation, we seek a way to leverage the error-bias property for better convergence of CNN model. To this end, we propose anisotropic direction loss (ADL) and anisotropic attention module (AAM) for coordinate and heatmap regression, respectively. ADL imposes strong binding force in normal direction for each landmark point on facial boundaries. On the other hand, AAM is an attention module which can get anisotropic attention mask focusing on the region of point and its local edge connected by adjacent points, it has a stronger response in tangent than in normal, which means relaxed constraints in the tangent. These two methods work in a complementary manner to learn both facial structures and texture details. Finally, we integrate them into an optimized end-to-end training pipeline named ADNet. Our ADNet achieves state-of-the-art results on 300W, WFLW and COFW datasets, which demonstrates the effectiveness and robustness.

1. Introduction

Face alignment, applied to facial landmark detection, has experienced tremendous improvement by means of Convolutional Neural Networks, and provides a continuous impetus for improvements in many computer vision techniques for face such as face recognition [29], face synthesis [1, 17] and face 3D reconstruction [23].

Error-bias can be treated as a special kind of AI-bias that could be resulted from the prejudiced assumptions made in the process of algorithm development or prejudices in the

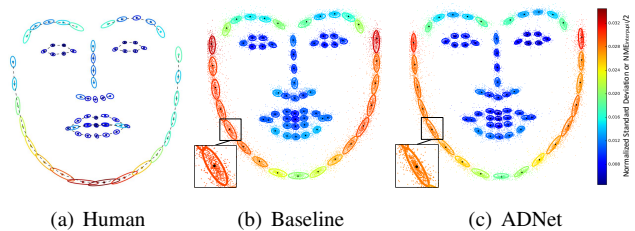---

*Corresponding author

Figure 1. Manual annotations variance v.s. Prediction error distribution on 300W dataset. (a) The variance of the manual annotations for each landmarks cited from 300W [36], the ellipses is colored by standard deviation normalized by face size. (b) The prediction error distribution of baseline model, each point represent the relevant position between predicted landmark and its corresponding ground truth, colored by the average $NME_{interpupil}$ / 2 of that landmark, in general face size is twice of the inter pupil. (c) The prediction error distribution of ADNet model.

training data thus is an anomaly in the output of machine learning algorithms. Common AI-bias, such as race-bias and gender-bias, always causes negative ethical effects, especially in the case of gender shades [5]. Different from them, the error-bias considered in this paper is more like a location uncertainty, as mentioned by [26].

In our research, error-bias is regarded as the nature of error direction and proves to be highly conducive for model understanding, thus deserving a thorough investigation to fill in the research gap.

Figure 1.(a) demonstrates the labeling-bias of landmarks location on 300W dataset, deduced by the anisotropic standard deviation of each point, especially the points located on the boundary. By the observation, we trained a common face alignment model with 300W dataset, whose setting is the same as the baseline model described at the end of Section 4.2, and rendered the error distribution on the whole test dataset in Figure 1.(b), in which the error means the relative offset from predicted position to ground truth. It is found that the error-bias exists on common face alignment model and reaches 33% relative rate, which is highly consistent with labeling-bias in Figure 1.(a). Based on the finding, we speculate that model easier converges the error

in normal than in tangent direction because of noisy label and semantic confusion in tangent direction. Inspired by this, in order to test the applicability, stronger and weaker constraint are respectively imposed in normal and tangent direction by the proposed ADNet instead of isotropic loss. Similar to Figure 1.(b), the corresponding error distribution is shown in Figure 1.(c), in which the error-bias becomes higher at 48%, but the error distribution becomes compact. To conclude, these three figures support our guideline that imposing stronger constraint to normal than to tangent.

In order to improve the localization capability of face alignment and fully leverage the error-bias, this paper proposes an end-to-end training framework involving devised Symmetric Direction Loss and Anisotropic Attention Module. Specifically, Anisotropic Direction Loss disentangles the landmark errors into normal error and tangent error, and imposes strong constraint in normal error and weak constraint in tangent error for coordinate regression, which is an improved L_n loss. Anisotropic Attention Module combines point heatmap and edge heatmap into one heatmap, which contains both landmarks information and local boundary information. Applying the combined attention heatmap as a mask to the landmarks heatmap, the model has high tolerance to tangent direction and low tolerance to normal direction. These two modules are highly aligned with the proposed guideline. Some previous works, such as LAB [47], PropNet [22], incorporates boundary information into CNN by attention and can be treated as special cases for the proposed guideline. ADNet enables both heatmap regression and coordinate regression optimization, and the anisotropic attention mask in it acts like a gabor filter supervised by both point and edge information.

The method is evaluated on several academic datasets, 300W [37], WFLW [47] and COFW [6]. All of them achieve the start-of-the-art performance, which demonstrates the effectiveness and robustness of the method.

In summary, the main contributions of this paper is as follows:

- Unveiling error-bias on error directions of face alignment, which is highly consistent with labeling-bias by human, strong in tangent direction and weak in normal direction, and based on this firstly proposing the guideline to leverage error-bias in face alignment by magnification instead of suppression.

- Devising Anisotropic Direction Loss to assign uneven loss weights to the disentangled normal and tangent error for each landmarks coordinate and magnify the error-bias by the proposed guideline.

- Proposing Anisotropic Attention Module to generate anisotropic attention mask for each landmark heatmap and magnify the error-bias again by the proposed guideline.

- Constructing an advanced end-to-end training pipeline and implementing extensive experiments on various datasets, the result outperforms other state-of-the-art methods.

2. Related Work

In the course of face alignment development, some classic approaches of face alignment were proposed in the 1990s, e.g. AAM [8, 38, 39, 30, 24], ASM [9, 10, 32] and cascade regression [14, 51]. In recent years, driven by vigorous development of Deep Convolutional Neural Network (DCNN), CNN-based face alignment methods have achieved state-of-the-art performance and have drawn close attention from researchers. Currently, the spotlight is centered on two main branches of face alignment, that are coordinate regression and heatmap regression.

Coordinate regression methods [40, 42, 43, 28, 52, 56] directly regress facial landmarks based on the input without postprocessing. To solve this problem, Zhang [55] involves multiple tasks, learning landmarks and facial attributes into the model concurrently. Then MDM [43] designs a pipeline to train the model from coarse to fine by focusing more on detailed local information. The basic combination, ResNet [18], DenseNet [20] with L1, L2, Smooth L1 or wing loss [15] are commonly used in this type of methods.

Heatmap regression methods [46, 13, 12, 33, 2] predict an intermediate heatmap for each landmark, and then the highest response point of each heatmap or near it is the final detected coordinate of landmark, where UNet [35] and stacked HG [33] are applied frequently. Adaptive wing loss [45] and focal wing loss [22] are proposed to balance the weight of easy sample and hard sample. Boundary information is introduced into face alignment by LAB [47], PropNet [22] and ACENet [21], which supplies more structure information and helps network to know which region should be paid more attention to. Usually, heatmap regression methods outperform coordinate regression methods with the effort of complex and tricky postprocessing.

Apart from the popular models, there are techniques that can further advance face alignment performance, such as CoordConv [27], Anti-aliased CNN [53], Multi-view CNN block [3] and attention mechanism [47]. Among them, CoordConv proves to be instrumental for coordinate transformation problem in the vision task, such as object detection and generative modeling. Anti-aliased CNN with a special pooling layer, has advantages in translation invariance or ranging degrees of translation dependence. The multi-view CNN block is proven to be beneficial for landmark localization on account of multiple receptive fields and the various scale of images those fields bring about. Furthermore, the attention mechanism is able to guide a CNN to study valuable features and focus on salient regions, thus gaining great popularity among scholars nowadays.

3. ADNet

We design a network that employs stacked 4 hour-glasses (HGs) [4] as backbone. In each HG structure, three heatmaps are generated, respectively corresponding to a *landmarks heatmap*, an *edge heatmap* and a *point heatmap*. All heatmaps contribute losses to model training in different ways. Based on these heatmaps, an *anisotropic attention mask* is generated from the point and edge heatmaps. The attention mask can then impose anisotropic supervision upon the landmarks training, where an anisotropic direction loss is applied to the predicted landmark coordinates that are formed through soft-argmax operation. We also leverage components such as coordconv [27] and anti-aliased blocks [53] to improve performance further, as is also proposed by [22]. We name this network the ADNet, with AD standing for anisotropic direction. The overview structure of ADNet is detailedly illustrated in Figure 2.

3.1. Anisotropic Direction Loss (ADL)

Before proposing the new loss, we would like to review the L_n loss in Equation 1. Two of its special cases are L_1 and L_2, which are commonly used in machine learning tasks, e.g. regression, detection and GAN.

$$L_n(p_i, \hat{p}_i) = |p_i - \hat{p}_i|^n \qquad (1)$$

where p_i and $\hat{p}_i$ are respectively the predicted value and ground truth of coordinate in ith landmarks. However, when utilizing $L1$ loss, model learns much noise instead of valid information from easy samples and when applying L_2 to outliers, model easily gets to gradient explosion. With the development of L_n, $Smooth\ L_1$ is proposed to solve these two problems by piecewise-defining as Equation 2.

$$SmoothL_1(p_i, \hat{p}_i) = \begin{cases} 0.5L_2(p_i, \hat{p}_i), & |p_i - \hat{p}_i| < 1 \\ L_1(p_i, \hat{p}_i) - 0.5, & otherwise \end{cases}$$
$$(2)$$

The L_n series losses treat the error isotropically, even if there is error-bias on error direction of model or labeling-bias on annotations, such as in landmarks regression. In other words, they ignore the anisotropy of errors in face alignment and give the same weight to loss in all directions.

According to the guideline of imposing strong constraint in normal direction and weak constraint in tangent direction, we propose the anisotropic direction loss (ADL), shown in Figure 3, which disentangles the error into two mutually orthogonal directions, namely normal error and tangent error, and put anisotropic loss weight to them, defined as follows:

$$ADL_n(p_i, \hat{p}_i) = \left(\frac{2\lambda}{1+\lambda} |N(p_i, \hat{p}_i) \cdot (p_i - \hat{p}_i)|^2 + \frac{2}{1+\lambda} |T(p_i, \hat{p}_i) \cdot (p_i - \hat{p}_i)|^2 \right)^{\frac{n}{2}}$$
$$(3)$$

Where $N(p_i, \hat{p}_i)$ and $T(p_i, \hat{p}_i)$ are unit vectors to disentangle error into two sub-errors, when the $\hat{p}_i$ locates on the edge, they are unit normal and tangent vectors respectively, otherwise, they would be unit error vectors; and λ refers to the hyper parameter to adjust constraint strength in these two sub-errors. Refer to the equation below for the corresponding $Smooth\ ADL_1$ loss:

$$SmoothADL_1(p_i, \hat{p}_i) =$$
$$\begin{cases} 0.5\ ADL_2(p_i, \hat{p}_i), & |p_i - \hat{p}_i| < 1 \\ ADL_1(p_i, \hat{p}_i) - 0.5, & otherwise \end{cases} \qquad (4)$$

The ADL_n loss can be treated as a more general form of the L_n loss. When $\lambda = 1$ or $\hat{p}_i$ does not subordinate to any edge, ADL_n exactly degenerates into L_n.

Generally, most landmarks locate on edges of either face contour or five sense organs. Hence, for these landmarks, we get the normal direction by calculating the slope from its directly adjacent points on the edge, namely $\hat{p}_{pre_i}$ and $\hat{p}_{next_i}$, which can be obtained from a pre-defined template of landmarks. And as for other landmarks not associating any edges, we still put isotropic constraint on all the directions, thus, both normal direction and tangent direction are equal to error direction. Based on this, normal and tangent direction can respectively be defined as below:

$$N(p_i, \hat{p}_i) = \begin{cases} \frac{\hat{p}_{pre_i} + \hat{p}_{next_i} - 2\hat{p}_i}{|\hat{p}_{pre_i} + \hat{p}_{next_i} - 2\hat{p}_i|}, & \hat{p}_i\ in\ edge \\ \frac{p_i - \hat{p}_i}{|p_i - \hat{p}_i|} & otherwise \end{cases} \qquad (5)$$

$$T(p_i, \hat{p}_i) = \begin{cases} S \times N(p_i, \hat{p}_i), & \hat{p}_i\ in\ edge \\ N(p_i, \hat{p}_i), & otherwise \end{cases} \qquad (6)$$

where S is a skew-symmetric matrix $\left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix} \right)$.

3.2. Anisotropic Attention Module (AAM)

As introduced in Section 2, coordinate regression and heatmap regression are widely applied in landmarks detection. Inspired by the proposed guideline, ADL loss is devised to deal with coordinate regression task under the bias of error direction. And as for heatmap regression task, in order to enhance the localization capability on normal direction, anisotropic attention module (AAM) is designed through supervising the network to focus more on tangent direction. In previous boundary-aware methods, the attention mask only contains boundary information, thus merely from which it is incapable of getting any point information.

To leverage the complete semantic information offered by prior information, the points and its directly connected edges, anisotropic attention module outputs the *anisotropic heatmap* by combining *point heatmap* and *edge heatmap* in Equation 7, named *point-edge heatmap*. When utilizing the *point-edge heatmap* as attention mask, ***the local region of***
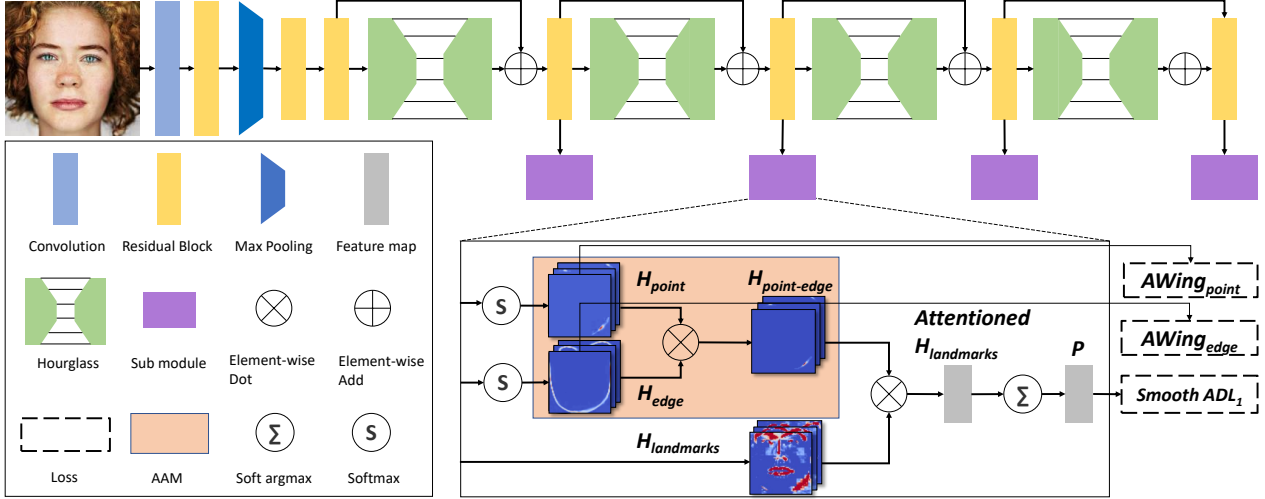
Figure 2. Overview of our ADL training framework.The backbone is constructed by stack four hourglass modules, and each hourglass module connected with three head branches, point attention, edge attention and final landmark regression branch.
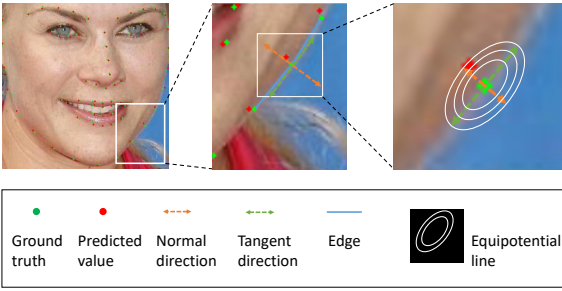

Figure 3. Diagram of Anisotropic Direction Loss (ADL). In the right figure, the green point ($\hat{p}_i$) attracts red point (p_i) under force filed, which is strong in normal direction and weak in tangent directions. According to the proposed guideline, the model studies more in normal direction.

target point with its partial edge has high response, which means it contains both edge and point information and can guide model training in normal and tangent anisotropically. Therefore, using *soft argmax* operation with the ability of mapping heatmap to coordinate, model accumulates landmarks coordinates mainly on the local region of attentioned landmarks heatmap with tangent information instead of on whole region of the landmarks heatmap.

$$H_{point-edge} = H_{point} \otimes H_{edge} \qquad (7)$$

Where H_{point} is the *point heatmap* and H_{edge} is the *edge heatmap*. Both of them are generated from the feature map of each HG and are supervised on two ground truth concurrently, point and edge heatmaps, by Awing [45] loss in Equation 8. Then the combined $H_{point-edge}$ serves as attention mask in landmarks coordinates regression by

Smooth ADL_1 loss in Equation 4. In this process, it is interesting that the shape of salient region in *point heatmap* H_{point} is more like a rough edge than a point region in gaussian distribution. This is caused by the supervisory signal from *Smooth* ADL_1 loss and is conducive to generate *point-edge heatmap* with strong tangent information.

$$AWing(y, \hat{y}) = \begin{cases} \omega \ln(1 + |\frac{y-\hat{y}}{\epsilon}|^{\alpha-y}), & |y - \hat{y}| < \theta \\ A|y - \hat{y}| - C, & otherwise \end{cases} \quad (8)$$

The detailed description and default values of hyper parameters could refer to AWing [45] loss, from which both H_{point} and H_{edge} adopt the same setting to learn the ground truth of respective heatmaps.

$$P = soft\ argmax(H_{landmarks} \otimes H_{point-edge}) \quad (9)$$

Where $H_{landmarks}$ is also one of the feature maps from each HG, and P is the predicted landmarks value by applying soft argmax operation to $H_{landmarks}$. Then network could learn P by annotated $\hat{P}$ for each training sample under the constraint of $Smooth\ ADL_1$.

Finally, the holistic loss of our architecture is:

$$\begin{aligned} L_{ADNet} = SmoothADL_1 \\ + \alpha \cdot AWing_{edge} + \beta \cdot AWing_{point} \end{aligned} \quad (10)$$

Where α and β are the loss weights to balance different tasks, we empirically set both as 10 in the paper.

4. Experiment

4.1. Implementation Details

To generate the inputs of our model, we cropped face regions and resized them into 256×256. We applied augmentation techniques to the training data by random rotation ($18°$), random scaling ($\pm 10\%$), random crop ($\pm 5\%$), random gray (20%), random blur (30%), random occlusion (40%) and random horizontal flip (50%). Regarding the architecture, we adopt four stacked hourglass modules. Each hourglass outputs three 64×64 feature maps (landmark, edge, and point heatmaps). To train our model, we empirically set the hyper-parameters of Gaussian sigma as 1.5 in point heatmap generation, line width as 1.0 in edge heatmap generation, and $\lambda = 2.0$ in the anisotropic direction loss function. We employed Adam optimizer with the initial learning rate of 1×10^{-3} and reduced the learning rate by 1/10 at each epoch of 200, 350, and 450. The model was trained on four GPUs (16GB NVIDIA Tesla P100), where the batch size of each GPU is 8.

4.2. Evaluation Metrics

Normalized Mean Error (NME) is a widely-used standard metric to evaluate landmark accuracy for face alignment algorithms, which is defined as

$$\text{NME}(P, \hat{P}) = \frac{1}{N_P} \sum_{i=1}^{N_P} \frac{\|p_i - \hat{p}_i\|_2}{d} \quad (11)$$

Where P and $\hat{P}$ denote the predicted and ground-truth coordinates of landmarks, respectively, N_P is the number of landmarks, and d is the reference distance to normalize the absolute errors. Usually, d could be inter-ocular (distance between outer eye corners) or inter-pupils (distance between pupil centers) distance. This paper uses inter-ocular distance as the normalization factor for 300W [37], WFLW [47], and inter-pupils for 300W [37] and COFW [6].

Failure Rate (FR) is a metric to evaluate the robustness of algorithms in terms of NME. Samples having larger NME than a pre-defined threshold are regarded as failed prediction. FR is defined by the percentage of failed examples over the whole dataset. We set the thresholds by 5% for 300W [37], and 10% for COFW [6] and WLFW [47].

Area Under Curve (AUC) is another widely-adopted metric for face alignment task. It can be calculated by using Cumulative Error Distribution (CED) curve. The horizontal axis of CED plot indicates the target NME (ranging between 0 and the pre-defined threshold), and the CED value at the specific point means the rate of samples whose NME are smaller than the specific target NME.

Method	NME	$FR_{10\%}$	$AUC_{10\%}$
Human [6]	5.60	-	-
RCPR [6]	8.50	20.00	-
TCDCN [54]	8.05	-	-
DAC-CSR [16]	6.03	4.73	-
Wu et al [49]	5.93	-	-
Wing [15]	5.44	3.75	-
DCFE [44]	5.27	7.29	0.3586
Awing [45]	4.94	0.99	**0.6440**
ADNet(Ours)	**4.68**	**0.59**	0.5317

Table 1. Comparing with state-of-the-art methods on COFW.

4.3. Method Comparison

To compare the performance of ADNet with the state-of-the-art competitors, we conduct evaluation on three public datasets: COFW [6], 300W [37] and WFLW [47].

COFW [6] contains a wide range of head poses and heavy occlusions, with 1,345 training images and 507 testing images. Each face has 29 annotated landmarks.

The experimental results on the COFW dataset are shown in Table 1, where all the models were tested under the same condition without any additional annotations. As tabulated in the table, our method outperforms the competitors by a wide margin in NME and $FR_{10\%}$. Compared with the second leading method, ADNet decreases NME by 5.3% in NME, which implies that the model is robust to large head poses and heavy occlusion despite the sparse landmarks definition.

300W [37] is a widely-adopted dataset for face alignment, with 3,148 images for training and 689 images for testing. The testing data is divided into two sub-categories: 554 for the common subset, and 135 for the challenging subset. All the data are manually labelled with 68 landmarks.

Table 2 compares our methods with the other approaches using different normalization factors: inter-ocular and inter-pupils distances. In both settings, our method yields the competitive accuracy, usually one of the best two. In particular, for inter-ocular normalization, ADNet improves the metric by 4.6% and 7.0% in NME over the previous SOTA method on the fullset and the common subset, respectively. As shown in Table 2, ADNet achieves a larger improvement on the common subset than the challenging subset, which could be due to the inaccurate labels in the challenging subset. For instance, in the challenging subset of 300W, we observed some samples having inaccurate landmark labels on the chin regions due to wrong bounding boxes, which could lead to the domain gap issue and degraded performance.

WFLW [47] is a challenging dataset constructed from in-the-wild face images. It also provides rich attribute labels such as occlusion, pose, make-up, illumination, blur and

Method	Common Subset	Challenging Subset	Fullset
Inter-pupil Normalization			
CFAN [52]	5.50	16.78	7.69
SDM [51]	5.57	15.40	7.50
LBF [34]	4.95	11.98	6.32
CFSS [57]	4.73	9.98	5.76
TCDCN [55]	4.80	8.60	5.54
MDM [43]	4.83	10.14	5.88
RAR [50]	4.12	8.35	4.94
DVLN [48]	3.94	7.62	4.66
TSR [28]	4.36	7.56	4.99
DSRN [31]	4.12	9.68	5.21
RCN$_+$(L+ELT) [19]	4.20	7.78	4.90
DCFE [44]	3.83	7.54	4.55
LAB [47]	3.42	6.98	4.12
Wing [15]	**3.27**	7.18	**4.04**
AWing [45]	3.77	6.52	4.31
ADNet(Ours)	3.51	**6.47**	4.08
Inter-ocular Normalisation			
PCD-CNN [25]	3.67	7.62	4.44
CPM+SBR [13]	3.28	7.58	4.10
SAN [13]	3.34	6.60	3.98
LAB [47]	2.98	5.19	3.49
DeCaFA [11]	2.93	5.26	3.39
DU-Net [41]	2.90	5.15	3.35
LUVLi [26]	2.76	5.16	3.23
AWing [45]	2.72	**4.52**	3.07
ADNet(Ours)	**2.53**	4.58	**2.93**

Table 2. Comparing with state-of-the-art methods on 300W.

expression. And it consists of 7,500 faces for training and 2,500 faces for testing, with 98 annotated landmarks.

Using the WFLW dataset, we aim to evaluate the models in various specific scenarios by making use of the annotated labels such as pose, expression, illumination and occlusion. In addition, in order to demonstrate the stability of landmark localization, we employ three metrics: NME, FR (10%) and AUC (10%). As tabulated in Table 3, ADNet outperforms the other state-of-the-art methods significantly in most of subsets.

4.4. Ablation study

To explore the efficacy and the contribution of each component of our proposed model, we conduct comprehensive ablation experiments.

Evaluation on different λ for ADL. To find the optimal hyper-parameters for ADL, we trained our model with different λ on the 300W dataset. And AAM is not included so as to avoid ambiguity. Instead of seeking the entire parameter space, we choose a few candidates of λ as shown in
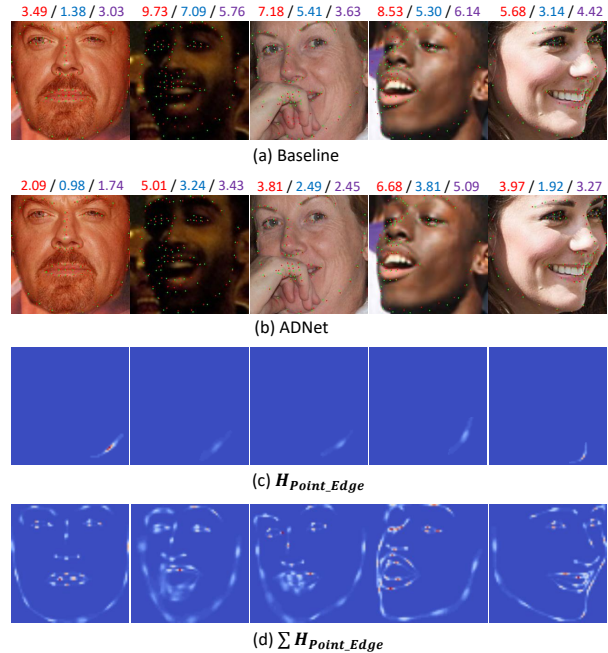

Figure 4. Samples from 300W fullset. (Red indicates NME, blue indicates normal NME, and purple indicates tangent NME.) (a) is the baseline result without ADL and AAM, (b) is the result of AD-Net, (c) is the output of a sample from $H_{point-edge}$ and (d) is the accumulated output of half $H_{point-edge}$ set for better visualization.

Table 4. It appears that, compared with the model without ADL ($\lambda = 1.0$), the optimal model gained about 4% improvement in NME when $\lambda = 2.0$. In addition, when λ is set to 0.5 (more weight for the tangent direction), the alignment accuracy dropped, which aligns with our assumption about error bias in face alignment. However, when $\lambda \geq 2.0$, the performance is not sensitive to the value of λ. In our experiments, we chose $\lambda = 2$ as the default setting.

Evaluation on different attention modules for AAM. AAM plays a vital role in ADNet, which contributes the largest improvement in alignment accuracy. In Table 5, 'w/ PAM' and 'w/ EAM' indicate that AAM is replaced with point attention module (PAM) and edge attention module (EAM), respectively. As revealed in the table, anisotropic attention leads to the best performance (12% improvement in NME over the second leading method) among the three ones that attempt to make the model focus on salient regions. As shown in Figure 4, the distribution of generated heatmap from AAM is aligned with the tangent direction rather than normal direction, which accounts for the effectiveness of AAM in our method.

Evaluation on different backbones in our method. To investigate the impact of backbones on face alignment per-

Metric	Method	Testset	Pose Subset	Expression Subset	Illumination Subset	Make-up Subset	Occlusion Subset	Blur Subset
$NME(\%)$	ESR [7]	11.13	25.88	11.47	10.49	11.05	13.75	12.20
	SDM [51]	10.29	24.10	11.45	9.32	9.38	13.03	11.28
	CFSS [57]	9.07	21.36	10.09	8.30	8.74	11.76	9.96
	DVLN [48]	6.08	11.54	6.78	5.73	5.98	7.33	6.88
	LAB [47]	5.27	10.24	5.51	5.23	5.15	6.79	6.12
	Wing [15]	5.11	8.75	5.36	4.93	5.41	6.37	5.81
	DeCaFA [11]	4.62	8.11	4.65	4.41	4.63	5.74	5.38
	AWing [45]	4.36	7.38	4.58	4.32	4.27	5.19	4.96
	LUVLi [26]	4.37	-	-	-	-	-	-
	ADNet(Ours)	**4.14**	**6.96**	**4.38**	**4.09**	**4.05**	**5.06**	**4.79**
	PropNet	4.05	6.92	**3.87**	4.07	3.76	4.58	4.36
	ADNet*(Ours)	**3.98**	**6.56**	4.02	**3.87**	**3.62**	**4.36**	**4.21**
$FR_{10\%}$	ESR [7]	35.24	90.18	42.04	30.80	38.84	47.28	41.40
	SDM [51]	29.40	84.36	33.44	26.22	27.67	41.85	35.32
	CFSS [57]	20.56	66.26	23.25	17.34	21.84	32.88	23.67
	DVLN [48]	10.84	46.93	11.15	7.31	11.65	16.30	13.71
	LAB [47]	7.56	28.83	6.37	6.73	7.77	13.72	10.74
	Wing [15]	6.00	22.70	4.78	4.30	7.77	12.50	7.76
	DeCaFA [11]	4.84	21.40	3.73	3.22	6.15	9.26	6.61
	AWing [45]	2.84	13.50	2.23	2.58	2.91	5.98	3.75
	LUVLi [26]	3.12	-	-	-	-	-	-
	ADNet(Ours)	**2.72**	**12.72**	**2.15**	**2.44**	**1.94**	**5.79**	**3.54**
	PropNet	2.96	12.58	2.55	2.44	1.46	5.16	3.75
	ADNet*(Ours)	**2.00**	**9.20**	**1.59**	**1.72**	**1.26**	**4.48**	**2.59**
$AUC_{10\%}$	ESR [7]	0.2774	0.0177	0.1981	0.2953	0.2485	0.1946	0.2204
	SDM [51]	0.3002	0.0226	0.2293	0.3237	0.3125	0.2060	0.2398
	CFSS [57]	0.3659	0.0632	0.3157	0.3854	0.3691	0.2688	0.3037
	DVLN [48]	0.4551	0.1474	0.3889	0.4743	0.4494	0.3794	0.3973
	LAB [47]	0.5323	0.2345	0.4951	0.5433	0.5394	0.4490	0.4630
	Wing [15]	0.5504	0.3100	0.4959	0.5408	0.5582	0.4885	0.4918
	DeCaFA [11]	0.5630	0.2920	**0.5460**	0.5790	0.5750	0.4850	0.4940
	AWing [45]	0.5719	0.3120	0.5149	0.5777	0.5715	0.5022	0.5120
	LUVLi [26]	0.5770	-	-	-	-	-	-
	ADNet(Ours)	**0.6022**	**0.3441**	0.5234	**0.5805**	**0.6007**	**0.5295**	**0.5480**
	PropNet	0.6158	0.3823	**0.6281**	0.6164	0.6389	0.5721	0.5836
	ADNet*(Ours)	**0.6250**	**0.4036**	0.6014	**0.6295**	**0.6406**	**0.5896**	**0.5903**

Table 3. Comparing with state-of-the-art methods on WFLW testing set. PropNet and ADNet*(Ours) employ focal wing loss [22] by using the attribute labels provided by WFLW.

Method	NME(%)	$FR_{5\%}$	$AUC_{5\%}$
$\lambda=0.5$	3.43	12.01	0.3504
$\lambda=1.0$(w/o ADL)	3.38	11.72	0.3653
$\lambda=2.0$	**3.23**	**10.45**	0.4006
$\lambda=4.0$	3.24	11.03	**0.4027**
$\lambda=8.0$	3.27	11.47	0.3950

Table 4. The contribution of ADL under different λ. AAM is not used in these settings.

Method	NME(%)	$FR_{5\%}$	$AUC_{5\%}$
w/o AAM	3.38	11.72	0.3653
w/ PAM	3.12	9.56	0.4095
w/ EAM	3.09	9.13	0.4124
w/ AAM	**2.98**	**8.72**	**0.4318**

Table 5. The contribution of AAM under different attention modules. ADL is not used in these settings.

formance, we trained models with two additional architectures. All the settings are identical to ADNet except for the network architecture. The results are tabulated in Table 6 where baseline denotes the model trained without ADL and AAM, in other words, baseline model feed

$H_{landmarks}$ directly to soft-argmax, then apply $Smooth\ L_1$ (a.k.a $Smooth\ ADL_1$ with $\lambda = 1$) loss to supervise landmark coordinates. For the ResNet50 backbone, we add three deconvolutional layers after the last layer to generate 64×64 heatmap like our original model. As shown in the table, all the models are significantly enhanced by equipping them with ADL and AAM, which demonstrates that ADL and AAM are generally adaptable regardless of network architecture. In addition, ADNet with the default backbone (Stacked 4 HGNet) yields the best performance among three models.

Backbone	Baseline	ADNet
Stacked 4 HGNet [33]	3.38	**2.93**
Single HGNet [33]	3.39	**2.99**
ResNet50 [18]	3.43	**3.11**

Table 6. NME(%) in different backbones on the 300W dataset.

Evaluation with and without facial attributes labels. It is shown that additional labeling information, such as face attributes, can help the model learn face alignment effectively. For example, PropNet [22] proposed a focal factor which dynamically adjusts the loss weight by utilizing class labels. We also test ADNet equipped with the focal factor, which is denoted as ADNet*. As shown in the Table 3 and 7, it is apparent that both PropNet [22] and ADNet* benefit from facial attribute labels.

Method	Tags	WFLW	300W
PropNet	w/	4.05	2.93
ADNet(Ours)	w/o	4.14	**2.93**
ADNet*(Ours)	w/	**3.98**	-

Table 7. NME(%) with and without attribute labels. The result of ADNet* on 300W is missing because the annotations for 300W was labelled by PropNet and the data is not published.

Evaluation on different error directions. Our fundamental assumption regarding the error-bias is that we should impose a strong constraint to the error along normal direction, while a relaxed constraint for the tangent-direction error. Similar ideas were also discussed in previous studies, LAB [47] and PropNet [22]. To further investigate the error direction, we report NME along normal and tangent directions independently as shown in Table 8 The baseline indicates the model without ADL and AAM. It can be observed that there is the significantly stronger NME bias along the tangent direction than tangent direction, which supports our assumption. In addition, ADL and AAM appear to conduce to model accuracy enhancement, especially in normal direction, while allowing flexibility along the tangent direction to some degree, resulting in the increased bias rate of 48.05%. In other words, it is reasonable to impose a strong

Method	Normal NME	Tangent NME	Overall NME	Bias rate
Baseline	1.91	2.55	3.38	33.51%
ADNet	**1.54**	**2.28**	**2.93**	**48.05%**

Table 8. NME(%) in normal and tangent directions of landmarks on the 300W dataset.

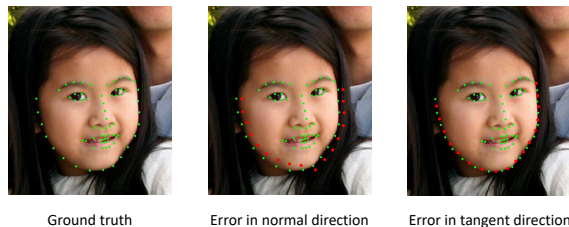Ground truth Error in normal direction Error in tangent direction

Figure 5. Visualization results on different error direction. Green points are ground truth, red points are mocked predictions. In the middle figure, all the mocked predictions have error in normal direction, and tangent direction in the right figure. Obviously, errors in tangent direction is more acceptable than normal direction which is aligned with labeling bias and human perception.

constraint to the error along the normal direction, which allows us to achieve better holistic performance in the end. In addition, as demonstrated in Figure 5, errors along the tangent direction (the third image) are more acceptable to human perception than the same amount of the errors along the normal direction (the second image).

$$Bias\ Rate = \frac{NME_{tangent} - NME_{normal}}{NME_{normal}} \quad (12)$$

Where $NME_{tangent}$ and NME_{normal} are respectively the NME in tangent and normal directions.

5. Conclusion

In this paper, we point out the significant bias of error distribution of each facial landmark, which is caused by ambiguity in the process of labeling. Inspired by this knowledge, anisotropic direction loss is proposed to handle error in normal and tangent directions independently. Moreover, anisotropic attention module is put forward to efficiently combine point and edge heatmap information. By combining and integrating the two modules, the proposed model imposes more constraint on normal direction than on tangent direction for the learning of landmarks coordinates. And rigorous experiments show that the proposed method outperforms other state-of-the-art models on multiple datasets. Finally, ablation studies verify the usefulness of the method comprehensively. Additionally, we realized that the current metrics do not reflect the actual error bias issue, and result in the same error values for the 2nd and 3rd images in Figure 5. In future work, the new metric considering human perception shown in Figure 5 and the error-bias should be studied to provide more meaningful insights.

References

[1] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Towards open-set identity preserving face synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6713–6722, 2018. 1

[2] Adrian Bulat and Georgios Tzimiropoulos. Convolutional aggregation of local evidence for large pose face alignment. 2016. 2

[3] Adrian Bulat and Georgios Tzimiropoulos. Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3706–3714, 2017. 2

[4] 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 *Proceedings of the IEEE International Conference on Computer Vision*, pages 1021–1030, 2017. 3

[5] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018. 1

[6] Xavier P Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust face landmark estimation under occlusion. In *Proceedings of the IEEE international conference on computer vision*, pages 1513–1520, 2013. 2, 5

[7] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2):177–190, 2014. 7

[8] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6):681–685, 2001. 2

[9] Timothy F Cootes and Christopher J Taylor. Active shape models—'smart snakes'. In *BMVC92*, pages 266–275. Springer, 1992. 2

[10] Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995. 2

[11] Arnaud Dapogny, Kevin Bailly, and Matthieu Cord. Decafa: deep convolutional cascade for face alignment in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6893–6901, 2019. 6, 7

[12] Jiankang Deng, George Trigeorgis, Yuxiang Zhou, and Stefanos Zafeiriou. Joint multi-view face alignment in the wild. *IEEE Transactions on Image Processing*, 28(7):3636–3648, 2019. 2

[13] Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Style aggregated network for facial landmark detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 379–388, 2018. 2, 6

[14] Zhen-Hua Feng, Guosheng Hu, Josef Kittler, William Christmas, and Xiao-Jun Wu. Cascaded collaborative regression for robust facial landmark detection trained using a mixture of synthetic and real images with dynamic weighting.

[15] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2235–2245, 2018. 2, 5, 6, 7

[16] Zhen-Hua Feng, Josef Kittler, William Christmas, Patrik Huber, and Xiao-Jun Wu. Dynamic attention-controlled cascaded shape regression exploiting training data augmentation and fuzzy-set sample weighting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2481–2490, 2017. 5

[17] Shuyang Gu, Jianmin Bao, Hao Yang, Dong Chen, Fang Wen, and Lu Yuan. Mask-guided portrait editing with conditional gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3436–3445, 2019. 1

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 8

[19] Sina Honari, Pavlo Molchanov, Stephen Tyree, Pascal Vincent, Christopher Pal, and Jan Kautz. Improving landmark localization with semi-supervised learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1546–1555, 2018. 6

[20] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 2

[21] Jihua Huang and Amir Tamrakar. Ace-net: Fine-level face alignment through anchors and contours estimation. *arXiv preprint arXiv:2012.01461*, 2020. 2

[22] Xiehe Huang, Weihong Deng, Haifeng Shen, Xiubao Zhang, and Jieping Ye. Propagationnet: Propagate points to curve to learn structure information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7265–7274, 2020. 2, 3, 7, 8

[23] Dalong Jiang, Yuxiao Hu, Shuicheng Yan, Lei Zhang, Hongjiang Zhang, and Wen Gao. Efficient 3d reconstruction for face recognition. *Pattern Recognition*, 38(6):787–798, 2005. 1

[24] Fatih Kahraman, Muhittin Gokmen, Sune Darkner, and Rasmus Larsen. An active illumination and appearance (aia) model for face alignment. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7. IEEE, 2007. 2

[25] Amit Kumar and Rama Chellappa. Disentangling 3d pose in a dendritic cnn for unconstrained 2d face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 430–439, 2018. 6

[26] 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 *Proceedings of the IEEE/CVF Conference on Com-*

IEEE Transactions on Image Processing, 24(11):3425–3440, 2015. 2

puter Vision and Pattern Recognition, pages 8236–8246, 2020. 1, 6, 7

[27] Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. *arXiv preprint arXiv:1807.03247*, 2018. 2, 3

[28] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3317–3326, 2017. 2, 6

[29] Iacopo Masi, Yue Wu, Tal Hassner, and Prem Natarajan. Deep face recognition: A survey. In *2018 31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI)*, pages 471–478. IEEE, 2018. 1

[30] Iain Matthews and Simon Baker. Active appearance models revisited. *International journal of computer vision*, 60(2):135–164, 2004. 2

[31] Xin Miao, Xiantong Zhen, Xianglong Liu, Cheng Deng, Vassilis Athitsos, and Heng Huang. Direct shape regression networks for end-to-end face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5040–5049, 2018. 6

[32] Stephen Milborrow and Fred Nicolls. Locating facial features with an extended active shape model. In *European conference on computer vision*, pages 504–513. Springer, 2008. 2

[33] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016. 2, 8

[34] Shaoqing Ren, Xudong Cao, Yichen Wei, and Jian Sun. Face alignment at 3000 fps via regressing local binary features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1685–1692, 2014. 6

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2

[36] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: Database and results. *Image and vision computing*, 47:3–18, 2016. 1

[37] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 397–403, 2013. 2, 5

[38] Jason Saragih and Roland Goecke. A nonlinear discriminative approach to aam fitting. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007. 2

[39] Patrick Sauer, Timothy F Cootes, and Christopher J Taylor. Accurate regression procedures for active appearance models. In *BMVC*, pages 1–11, 2011. 2

[40] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3476–3483, 2013. 2

[41] Zhiqiang Tang, Xi Peng, Shijie Geng, Lingfei Wu, Shaoting Zhang, and Dimitris Metaxas. Quantized densely connected u-nets for efficient landmark localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 339–354, 2018. 6

[42] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660, 2014. 2

[43] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4177–4187, 2016. 2, 6

[44] Roberto Valle, Jose M Buenaposada, Antonio Valdes, and Luis Baumela. A deeply-initialized coarse-to-fine ensemble of regression trees for face alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 585–601, 2018. 5, 6

[45] Xinyao Wang, Liefeng Bo, and Li Fuxin. Adaptive wing loss for robust face alignment via heatmap regression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6971–6981, 2019. 2, 4, 5, 6, 7

[46] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016. 2

[47] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2129–2138, 2018. 2, 5, 6, 7, 8

[48] Wenyan Wu and Shuo Yang. Leveraging intra and inter-dataset variations for robust face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 150–159, 2017. 6, 7

[49] Yue Wu and Qiang Ji. Robust facial landmark detection under significant head poses and occlusion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3658–3666, 2015. 5

[50] Shengtao Xiao, Jiashi Feng, Junliang Xing, Hanjiang Lai, Shuicheng Yan, and Ashraf Kassim. Robust facial landmark detection via recurrent attentive-refinement networks. In *European conference on computer vision*, pages 57–72. Springer, 2016. 6

[51] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 532–539, 2013. 2, 6, 7

[52] Jie Zhang, Shiguang Shan, Meina Kan, and Xilin Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *European conference on computer vision*, pages 1–16. Springer, 2014. 2, 6

[53] Richard Zhang. Making convolutional networks shift-invariant again. In *International Conference on Machine Learning*, pages 7324–7334. PMLR, 2019. 2, 3

[54] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *European conference on computer vision*, pages 94–108. Springer, 2014. 5

[55] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Learning deep representation for face alignment with auxiliary attributes. *IEEE transactions on pattern analysis and machine intelligence*, 38(5):918–930, 2015. 2, 6

[56] Erjin Zhou, Haoqiang Fan, Zhimin Cao, Yuning Jiang, and Qi Yin. Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 386–391, 2013. 2

[57] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang. Face alignment by coarse-to-fine shape searching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4998–5006, 2015. 6, 7