Look at Boundary: A Boundary-Aware Face Alignment Algorithm

Wayne Wu [*1,2], Chen Qian[2], Shuo Yang[3], Quan Wang[2], Yici Cai[1], Qiang Zhou[1]

[1]Tsinghua National Laboratory for Information Science and Technology (TNList),
Department of Computer Science and Technology, Tsinghua University
[2]SenseTime Research
[3]Amazon Rekognition

[1]`wwy15@mails.tsinghua.edu.cn` [1]`caiyc@mail.tsinghua.edu.cn` [1]`zhouqiang@tsinghua.edu.cn`
[2]`{qianchen, wangquan}@sensetime.com` [3]`shuoy@amazon.com`

Abstract

We present a novel boundary-aware face alignment algorithm by utilising boundary lines as the geometric structure of a human face to help facial landmark localisation. Unlike the conventional heatmap based method and regression based method, our approach derives face landmarks from boundary lines which remove the ambiguities in the landmark definition. Three questions are explored and answered by this work: 1. Why using boundary? 2. How to use boundary? 3. What is the relationship between boundary estimation and landmarks localisation? Our boundary-aware face alignment algorithm achieves 3.49% mean error on 300-W Fullset, which outperforms state-of-the-art methods by a large margin. Our method can also easily integrate information from other datasets. By utilising boundary information of 300-W dataset, our method achieves 3.92% mean error with 0.39% failure rate on COFW dataset, and 1.25% mean error on AFLW-Full dataset. Moreover, we propose a new dataset WFLW to unify training and testing across different factors, including poses, expressions, illuminations, makeups, occlusions, and blurriness. Dataset and model will be publicly available at `https://wywu.github.io/projects/LAB/LAB.html`

1. Introduction

Face alignment, which refers to facial landmark detection in this work, serves as a key step for many face applications, e.g., face recognition [75], face verification [48, 49] and face frontalisation [21]. The objective of this paper is to devise an effective face alignment algorithm to handle faces with unconstrained pose variation and occlusion across multiple datasets and annotation protocols.

---

[*]This work was done during an internship at SenseTime Research.
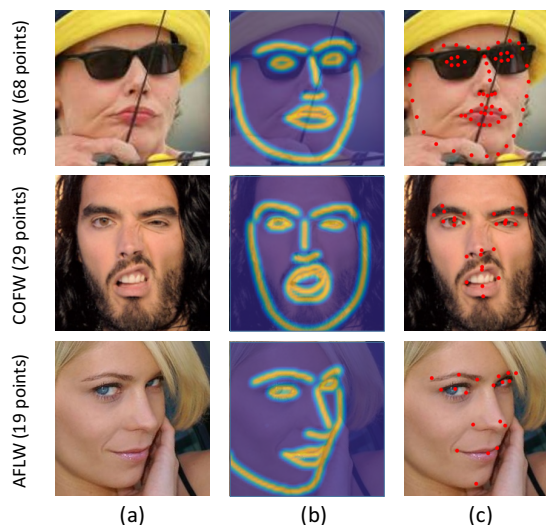

Figure 1: The first column shows the face images from different datasets with different number of landmarks. The second column illustrates the universally defined facial boundaries estimated by our methods. With the help of boundary information, our approach achieves high accuracy localisation results across multiple datasets and annotation protocols, as shown in the third column.

Different to face detection [45] and recognition [75], face alignment identifies geometry structure of human face which can be viewed as modeling highly structured output. Each facial landmark is strongly associated with a well-defined facial boundary, e.g., eyelid and nose bridge. However, compared to boundaries, facial landmarks are not so well-defined. Facial landmarks other than corners can hardly remain the same semantic locations with large pose variation and occlusion. Besides, different annotation schemes of existing datasets lead to a different number of landmarks [28, 5, 66, 30] (19/29/68/194 points) and annotation scheme of future face alignment datasets can hardly be determined. We believe the reasoning of a unique facial

structure is the key to localise facial landmarks since human face does not include ambiguities.

To this end, we use well-defined facial boundaries to represent the geometric structure of the human face. It is easier to identify facial boundaries comparing to facial landmarks under large pose and occlusion. In this work, we represent facial structure using 13 boundary lines. Each facial boundary line can be interpolated from a sufficient number of facial landmarks across multiple datasets, which will not suffer from inconsistency of the annotation schemes.

Our boundary-aware face alignment algorithm contains two stages. We first estimate facial boundary heatmaps and then regress landmarks with the help of boundary heatmaps. As noticed in Fig. 1, facial landmarks of different annotation schemes can be derived from boundary heatmaps with the same definition. To explore the relationship between facial boundaries and landmarks, we introduce adversarial learning ideas by using a landmark-based boundary effectiveness discriminator. Experiments have shown that the better quality estimated boundaries have, the more accurate landmarks will be. The boundary heatmap estimator, landmark regressor, and boundary effectiveness discriminator can be jointly learned in an end-to-end manner.

We used stacked hourglass structure [35] to estimate facial boundary heatmap and model the structure between facial boundaries through message passing [11, 63] to increase its robustness to occlusion. After generating facial boundary heatmaps, the next step is deriving facial landmarks using boundaries. The boundary heatmaps serve as structure cue to guide feature learning for the landmark regressor. We observe that a model guided by ground truth boundary heatmaps can achieve 76.26% AUC on 300W [39] test while the state-of-the-art method [15] can only achieve 54.85%. This suggests the richness of information contained in boundary heatmaps. To fully utilise the structure information, we apply boundary heatmaps at multiple stages in the landmark regression network. Our experiment shows that the more stages boundary heatmaps are used in feature learning, the better landmark prediction results we will get.

We evaluate the proposed method on three popular face alignment benchmarks including 300W [39], COFW [5], and AFLW [28]. Our approach significantly outperforms previous state-of-the-art methods by a large margin. 3.49% mean error on 300-W Fullset, 3.92% mean error with 0.39% failure rate on COFW and 1.25% mean error on AFLW-Full dataset respectively. To unify the evaluation, we propose a new large dataset named Wider Facial Landmarks in-the-wild (WFLW) which contain 10,000 images. Our new dataset introduces large pose, expression, and occlusion variance. Each image is annotated with 98 landmarks and 6 attributes. Comprehensive ablation study demonstrates the effectiveness of each component.

2. Related Work

In the literature of face alignment, besides classic methods (ASMs [34, 23], AAMs [13, 41, 33, 25], CLMs [29, 42] and Cascaded Regression Models [7, 5, 58, 8, 72, 73, 18]), recently, state-of-the-art performance has been achieved with Deep Convolutional Neural Networks (DC-NNs). These methods mainly fall into two categories, *i.e.*, coordinate regression model and heatmap regression model.

Coordinate regression models directly learn the mapping from the input image to the landmark coordinates vector. Zhang *et al.* [70] frames the problem as a multi-task learning problem, learns landmark coordinates and predicts facial attributes at the same time. MDM [51] is the first end-to-end recurrent convolutional system for face alignment from coarse to fine. TSR [31] splits face into several parts to ease the parts variations and regresses the coordinates of different parts respectively. Even though coordinate regression models have the advantage of explicit inference of landmark coordinates without any post-processing. Nevertheless, they are not performing as well as heatmap regression models.

Heatmap regression models, which generate likelihood heatmaps for each landmark respectively, have recently achieved state-of-the-art performance in face alignment. CALE [4] is a two-stage convolutional aggregation model to aggregate score maps predicted by detection stage along with early CNN features for final heatmap regression. Yang *et al.* [60] uses a two parts network, *i.e.*, a supervised transformation to normalise faces and a stacked hourglass network [35] to get prediction heatmaps. Most recently, JMFA [15] achieves state-of-the-art accuracy by leveraging stacked hourglass network [35] for multi-view face alignment and demonstrates better than the best three entries of the last Menpo Challenge [66].

Since boundary detection was set as one of the most fundamental problems in computer vision and there have emerged a large number of materials [56, 52, 44, 65, 43]. It has been proved efficient in vision tasks as segmentation [32, 27, 22] and object detection [36, 50, 37]. In face alignment, boundary information demonstrates especial importance because almost all of the landmarks are defined lying on the facial boundaries. However, as far as we know, in face alignment task, no work before has investigated the use of boundary information from an *explicit* perspective.

The recent advance in human pose estimation partially inspires our method of boundary heatmaps estimation. Stacked hourglass network [35] achieves compelling accuracy with a bottom-up, top-down design which endows the network with capabilities of obtaining multi-scale information. Message passing [11, 63] has shown great power in structure modeling of human joints. Recently, adversarial learning [9, 10] is adopted to further improve the accuracy of estimated human pose under heavy occlusion.
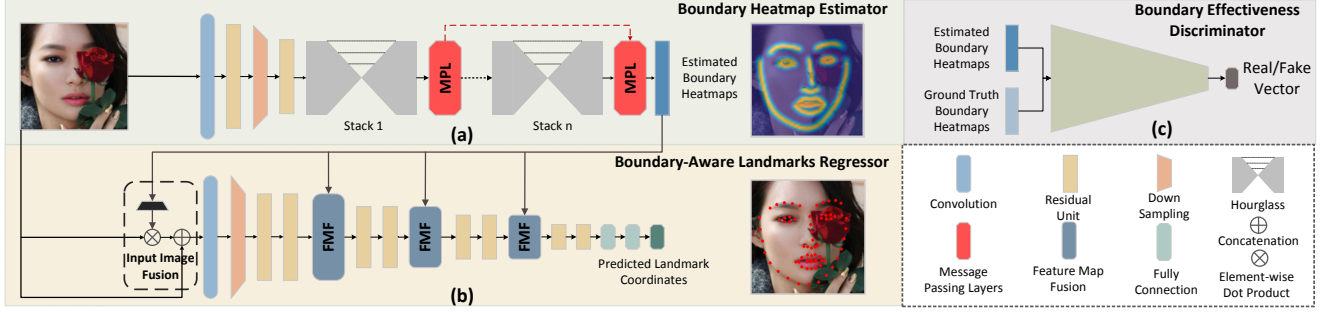
Figure 2: Overview of our Boundary-Aware Face Alignment framework. (a) Boundary heatmap estimator, which based on hourglass network is used to estimate boundary heatmaps. Message passing layers are introduced to handle occlusion. (b) Boundary-aware landmarks regressor is used to generate the final prediction of landmarks. Boundary heatmap fusion scheme is introduced to incorporate boundary information into the feature learning of regressor. (c) Boundary effectiveness discriminator, which distinguishes "real" boundary heatmaps from "fake", is used to further improve the quality of the estimated boundary heatmaps.

3. Boundary-Aware Face Alignment

As mentioned in the introduction, landmarks have difficulty in presenting accurate and universal geometric structure of face images. We propose facial boundary as geometric structure representation and help landmarks regression problem in the end. Boundaries are detailed and well-defined structure descriptions, which are consistent across head poses and datasets. They are also closely related to landmarks since most of the landmarks are located along boundary lines.

Other choices are also available for geometric structure representations. Recent works [31, 47, 19] has adopted facial parts to aid face alignment tasks. However, facial parts are too coarse thus not as powerful as boundary lines. Another choice would be face parsing results. Face parsing leads to disjoint facial components which needs the boundaries of each component form a closed loop. However, some facial organs such as nose are naturally blended into the whole face thus are inaccurate to be defined as separate parts. On the contrary, boundary lines are not necessary to form a closed loop, which is more flexible in representing geometric structure. Experiments in Sec 4.2 have shown that boundary lines are the best choice to aid landmark coordinates regression.

The detailed configuration of our proposed Boundary-Aware Face Alignment framework is illustrated in Fig. 2. It is composed of three closely related components: Boundary-Aware Landmark Regressor, Boundary Heatmap Estimator and Landmark-Based Boundary Effectiveness Discriminator. Boundary-Aware Landmark Regressor incorporates boundary information in a multi-stage manner to predict landmark coordinates. Boundary Heatmap Estimator produces boundary heatmaps as face geometric structure. Since boundary information is used heavily, the quality of boundary heatmaps is crucial for final landmark regression. We introduce adversarial learning idea [20] by proposing Landmark-Based Boundary Effectiveness Dis-

criminator, which is paired with the Boundary Heatmap Estimator. This discriminator can further improve the quality of boundary heatmaps and lead to better landmark coordinates prediction.

3.1. Boundary-aware landmarks regressor

In order to fuse boundary line into feature learning, we transform landmarks to boundary heatmaps to aid the learning of feature. The responses of each pixel in boundary heatmap are decided by its distance to the corresponding boundary line. As shown in Fig. 3, the details of boundary heatmap are defined as follows.

Given a face image I, denote its ground truth annotation by L landmarks as $S = \{s_l\}_{l=1}^{L}$. K subsets $S_i \subset S$ are defined to represent landmarks belongs to K boundaries respectively, such as upper left eyelid and nose bridge. For each boundary, S_i is interpolated to get a dense boundary line. Then a binary boundary map B_i, the same size as I, is formed by setting only points on the boundary line to be 1, others 0. Finally, a distance transform is performed based on each B_i to get distance map D_i. We use a gaussian expression with standard deviation σ to transform the distance map to ground-truth boundary heatmap M_i. 3σ is used to threshold D_i to make boundary heatmaps focus more on boundary areas. In practice, the length of the ground-truth boundary heatmap side is set to a quarter of the size of I for computation efficiency.

$$M_i(x,y) = \begin{cases} \exp(-\frac{D_i(x,y)^2}{2\sigma^2}), & \text{if } D_i(x,y) < 3\sigma \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

In order to fully utilise the rich information contained in boundary heatmaps, we propose a multi-stage boundary heatmap fusion scheme. As illustrated in Fig. 2, A four-stage res-18 network is adopted as our baseline network. Boundary heatmap fusion is conducted at the input and every stage of the network. Comprehensive results in Sec. 4.2
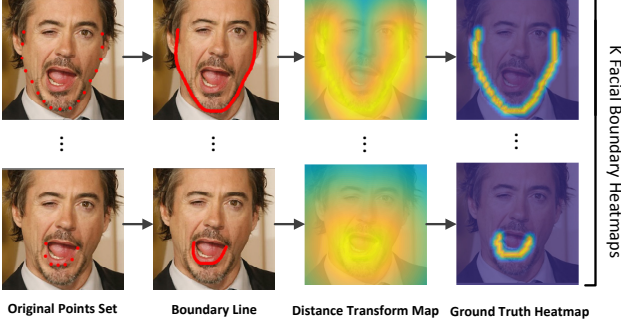
Figure 3: An illustration of the process of ground truth heatmap generation. Each row represents the process of one specific facial boundary, *i.e.*, facial outer contour, left eyebrow, right eyebrow, nose bridge, nose boundary, left/right upper/lower eyelid and upper/lower side of upper/lower lip.
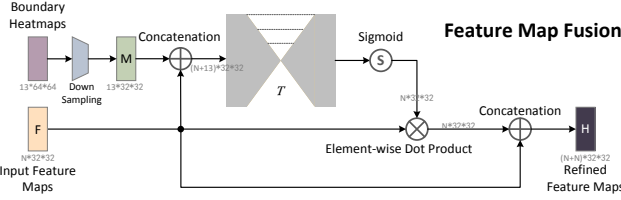

Figure 4: An illustration of the *feature map fusion* scheme. Boundary cues and input feature maps are fused together to get a refined feature with the usage of a hourglass module.

have shown that the more fusion we conducted to the baseline network, the better performance we can get.

Input image fusion. To fuse boundary heatmap M with input image I, the fused input H is defined as:

$$H = I \oplus (M_1 \otimes I) \oplus ... \oplus (M_T \otimes I) \qquad (2)$$

where $\otimes$ represents the element-wise dot product operation and $\oplus$ represents channel-wise concatenation. The above design makes fused input focus only on detailed texture around boundaries. Thus most background and texture-less face regions are ignored which greatly enhance the effectiveness of input. The original input is also concatenated to the fused ones to keep other valuable information in the original image.

Feature map fusion. Similar to above, to fuse boundary heatmap M with feature map F, the fused feature map H is defined as:

$$H = F \oplus (F \otimes T(M \oplus F)) \qquad (3)$$

Since the number of channels of M equals to the number of pre-defined boundaries, which is constant. A transform function T is necessary to convert M to have the same channels with F. We choose hourglass structure subnet as T to keep feature map size. Down-sampling and up-sampling are performed symmetrically. Skip connections
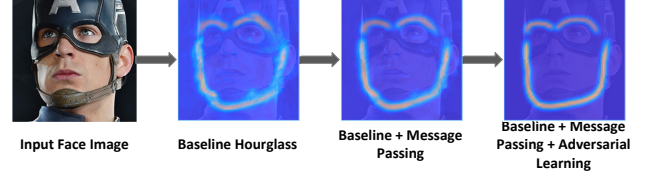

Figure 5: An illustration of the effectiveness of message passing and adversarial learning. With the message passing and adversarial learning addition, the quality of the estimated boundary is well improved to be more and more plausible and focused.

are used to combine multi-scale information. Then a sigmoid layer normalises the output range to $[0, 1]$. Another simple choice would be consecutive convolutional layers with stride equals to one, which covers relatively local areas. Experiments in Sec. 4.2 have demonstrated the superiority of hourglass structure. Details of feature map fusion subnet are illustrated in Fig. 4.

Since boundary heatmaps are used heavily in landmarks coordinates regression. The quality of boundary heatmaps is essential to the prediction accuracy. By fusing ground truth boundary heatmaps, our method can achieve 76.26% AUC on 300-W test, comparing to the state-of-art result 54.85%. Based on this experiment, in the following sections, several methods will be introduced to improve the quality of generated boundary heatmaps. Experiment in ablation study also shows the consistent performance gain with better heatmap quality.

3.2. Boundary heatmap estimator

Following previous work in face alignment [15, 60] and human pose [35, 12, 62], we use stacked hourglass as the baseline of boundary heatmap estimator. Mean square error (MSE) between generated and groundtruth boundary heatmaps is optimized. However, as demonstrated in Fig. 5, when heavy occlusions happen, the generated heatmaps always suffer from the noisy and multi-mode response, which has also been mentioned in [9, 12].

In order to relieve the problem caused by occlusion, we introduce message passing layers to pass information between boundaries. This process is visualised in Fig. 6. During occlusion, visible boundaries can provide help to occluded ones according to face structure. **Intra-level message passing** is used at the end of each stack to pass information between different boundary heatmaps. Thus, information can be passed from visible boundaries to occluded ones. Moreover, since different stacks of hourglass focus on different aspects of face information. **Inter-level message passing** is adopted to pass message from lower stacks to the higher stacks to keep the quality of boundary heatmaps when stacking more hourglass subnets.

We implemented message passing following [11]. In this implementation, the feature map at the end of each stack needs be divided into K branches, where K is the number
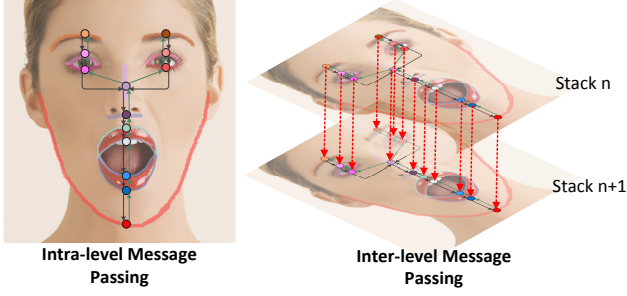
Figure 6: An illustration of message pass scheme. A bi-direction tree structure is used for intra-level message passing. Inter-level message is passed between adjacent stacks from lower to higher.

of boundaries, each represents a type of boundary feature map. This requirement demonstrates the advantage of our boundary heatmaps compared with landmark heatmaps [15, 60] for the small and constant number K of them. Thus, the computational and parameter cost of message passing layers within boundaries is small while it is not practical for message passing within 68 or even 194 landmarks.

3.3. Boundary effectiveness discriminator

In structured boundary heatmap estimator, mean squared error (MSE) is used as the loss function. However, minimizing MSE sometimes makes the prediction look blurry and implausible. This regression-to-the-mean problem is a well-known fact in the literature of super-resolution [40]. It damages the learning of regression network when bad boundary heatmaps are generated.

However, in our framework, the hard-to-define term "quality" of heatmaps has a very clear evaluation metric. If helping to produce accurate landmark coordinates, the boundary heatmap has a good quality. According to this, we propose a landmark based boundary effectiveness discriminator to decide the effectiveness of the generated boundary heatmaps. For a generated boundary heatmap $\hat{M}$ (all index i such as $\hat{M}_i$ is omitted for the simplicity of notation), denote its corresponding generated landmark coordinates set as $\hat{S}$, the ground-truth distance matric map as $Dist$. The ground truth d_{fake} of discriminator D that determines whether the generated boundary heatmap is fake can be defined as

$$d_{\text{fake}}(\hat{M}, \hat{S}) = \begin{cases} 0, & \Pr_{s \in \hat{S}}(Dist(s) < \theta) < \delta \\ 1, & \text{otherwise} \end{cases} \quad (4)$$

Where θ is the distance threshold to ground truth boundary and δ is the probability threshold. This discriminator predicts whether most generated corresponding landmarks would be close to the ground truth boundary.

Following [9, 10], we introduce the idea of adversarial learning by pairing the boundary effectiveness discriminator D and the boundary heatmaps estimator G. The loss of D can be expressed as:

$$\mathcal{L}_D = -(\mathbb{E}[\log D(M)] + \mathbb{E}[\log(1 - |D(G(I)) - d_{\text{fake}}|)]) \quad (5)$$

Where M is the ground truth boundary heatmap. The discriminator learns to predict ground truth boundary heatmap as one while predict generated boundary heatmap according to d_{fake}.

With effectiveness discriminator, the adversarial loss can be expressed as:

$$\mathcal{L}_A = \mathbb{E}[\log(1 - D(G(I)))] \quad (6)$$

Thus, the estimator is optimised to fool D by giving more plausible and high-confidence maps that will benefit the learning of regression network.

The following pseudo-code shows the training process of the whole methods.

Algorithm 1 The training pipeline of the our method.

Require: Training image I, the corresponding ground-truth boundary heatmaps M and landmark coordinates S, the generation network G, the regression network R and the discrimination network D.

1: **while** the accuracy of landmarks predicted by R in validation set stops **do**
2:      Forward G by $\hat{M} = G(I)$ and optimize G by minimizing $\|\hat{M} - M\|_2^2 + \mathcal{L}_A$ where $\mathcal{L}_A$ is defined in Eq.6;
3:      Forward D by $\hat{d}_{real} = D(M)$ and optimize D by minimizing the first term of $\mathcal{L}_D$ defined in Eq.5;
4:      Forward D by $\hat{d}_{fake} = D(\hat{M})$ and optimize D by minimizing the second term of $\mathcal{L}_D$ defined in Eq.5;
5:      Forward R by $\hat{S} = R(I, \hat{M})$ and optimize R by minimizing $\|\hat{S} - S\|_2^2$;
6: **end while**

3.4. Cross-Dataset Face Alignment

Recently, together with impressive progress of algorithms for face alignment, various benchmarks have also been released, *e.g.*, LFPW [3], AFLW [28] and 300-W [39]. However, because of the gap between annotation schemes, these datasets can hardly be jointly used. Models trained on one specific dataset perform poorly on recent in-the-wild test sets.

However, introduction of an annotation transfer component [46, 71, 67, 53] will bring new problems. From a new perspective, we take facial boundaries as an all-purpose middle-level face geometry representation. Facial boundaries naturally unify different landmark definitions with enough landmarks. And it can also be applied to help training landmarks regressor with any specific landmarks definition. The cross-dataset capacity is an important by-product of our methods. Its effectiveness is evaluated in Sec. 4.1.

4. Experiments

Datesets. We conduct evaluation on four challenging datasets including 300W [39], COFW [5], AFLW [28] and WFLW which is annotated by ourself.

300W [39] dataset: 300W is currently the most widely-used benchmark dataset. We regard all the training samples (3148 images) as the training set and perform testing on *(i)* full set and *(ii)* test set. *(i)* Full set contains 689 images and is split into common subset (554 images) and challenging subsets (135 images). *(ii)* Test set is the private test-set used for the 300W competition which contains 600 images.

COFW [5] dataset consists of 1345 images for training and 507 faces for testing which are all occluded to different degrees. Each COFW face originally has 29 manually annotated landmarks. We also use the test set which has been re-annotated by [19] with 68 landmarks annotation scheme to allow easy comparison to previous methods.

AFLW [28] dataset: AFLW contains 24386 in-the-wild faces with large head pose up to $120°$ for yaw and $90°$ for pitch and roll. We follow [72] to adopt three settings on our experiments: *(i)* AFLW-Full: 20000 and 4386 images are used for training and testing respectively. *(ii)* AFLW-Frontal: 1314 images are selected from 4386 testing images for evaluation on frontal faces.

WFLW dataset: In order to facilitate future research of face alignment, we introduce a new facial dataset base on WIDER Face [61] named Wider Facial Landmarks in-the-wild (WFLW), which contains 10000 faces (7500 for training and 2500 for testing) with 98 fully manual annotated landmarks. Apart from landmark annotation, out new dataset includes rich attribute annotations, i.e., occlusion, pose, make-up, illumination, blur and expression for comprehensive analysis of existing algorithms. Compare to previous dataset, faces in the proposed dataset introduce large variations in expression, pose and occlusion. We can simply evaluate the robustness of pose, occlusion, and expression on proposed dataset instead of switching between multiple evaluation protocols in different datasets. The comparison of WFLW with popular benchmarks is illustrated in the supplementary material.

Evaluation metric. We evaluate our algorithm using standard normalised landmarks mean error and Cumulative Errors Distribution (CED) curve. In addition, two further statistics *i.e.* the area-under-the-curve (AUC) and the failure rate for a maximum error of 0.1 are reported. Because of various profile face on AFLW [28] dataset, we follow [72] to use face size as the normalising factor. For other dataset, we follow MDM [51] and [39] to use outer-eye-corner distance as the "inter-ocular" normalising factor. Specially, to compare with the results that reported to be normalised by "inter-pupil" (eye-centre-distance) distance, we report our results with both two normalising factors on Table 1.

Implementation details. All training images are cropped

Method	Common Subset	Challenging Subset	Fullset
Inter-pupil Normalisation			
RCPR [6]	6.18	17.26	8.35
CFAN [69]	5.50	16.78	7.69
ESR [7]	5.28	17.00	7.58
SDM [57]	5.57	15.40	7.50
LBF [38]	4.95	11.98	6.32
CFSS [72]	4.73	9.98	5.76
3DDFA [74]	6.15	10.59	7.01
TCDCN [70]	4.80	8.60	5.54
MDM [51]	4.83	10.14	5.88
RAR [55]	4.12	8.35	4.94
DVLN [53]	3.94	7.62	4.66
TSR [31]	4.36	7.56	4.99
LAB	**3.42**	**6.98**	**4.12**
LAB+Oracle	2.57	4.72	2.99
Inter-ocular Normalisation			
PCD-CNN [2]	3.67	7.62	4.44
SAN [59]	3.34	6.60	3.98
LAB	**2.98**	**5.19**	**3.49**
LAB+Oracle	1.85	3.28	2.13

Table 1: Mean error (%) on 300-W Common Subset, Challenging Subset and Fullset (68 landmarks).

Method	AUC	Failure Rate (%)
Deng *et al*. [14]	0.4752	5.5
Fan *et al*. [16]	0.4802	14.83
DenseReg + MDM [1]	0.5219	3.67
JMFA [15]	0.5485	1.00
LAB	**0.5885**	**0.83**
LAB+Oracle	0.7626	0.00

Table 2: Mean error (%) on 300-W testset (68 landmarks). Accuracy is reported as the AUC and the Failure Rate.

and resized to 256×256 according to provided bounding boxes. The estimator is stacked four times if not specially indicated in our experiment. For ablation study, the estimator is stacked two times due to the consideration of time and computation cost. All our models are trained with *Caffe* [24] on 4 Titan X GPUs. Note that all testing images are cropped and resized according to provided bounding boxes *without any spatial transformation* for fair comparison with other methods. For the limited space of paper, we report all of the training details and experiment settings in our supplementary material.

4.1. Comparison with existing approaches

4.1.1 Evaluation on 300W

We compare our approach against the state-of-the-art methods on 300W Fullset. The results are shown in Table 1. Our method significantly outperforms previous methods by a large margin. Note that, our method achieves 6.98% mean

Metric	Method	Testset	Pose Subset	Expression Subset	Illumination Subset	Make-Up Subset	Occlusion Subset	Blur Subset
Mean Error (%)	ESR [7]	11.13	25.88	11.47	10.49	11.05	13.75	12.20
	SDM [57]	10.29	24.10	11.45	9.32	9.38	13.03	11.28
	CFSS [72]	9.07	21.36	10.09	8.30	8.74	11.76	9.96
	DVLN [53]	6.08	11.54	6.78	5.73	5.98	7.33	6.88
	LAB	**5.27**	**10.24**	**5.51**	**5.23**	**5.15**	**6.79**	**6.32**
Failure Rate (%)	ESR [7]	35.24	90.18	42.04	30.80	38.84	47.28	41.40
	SDM [57]	29.40	84.36	33.44	26.22	27.67	41.85	35.32
	CFSS [72]	20.56	66.26	23.25	17.34	21.84	32.88	23.67
	DVLN [53]	10.84	46.93	11.15	7.31	11.65	16.30	13.71
	LAB	**7.56**	**28.83**	**6.37**	**6.73**	**7.77**	**13.72**	**10.74**
AUC	ESR [7]	0.2774	0.0177	0.1981	0.2953	0.2485	0.1946	0.2204
	SDM [57]	0.3002	0.0226	0.2293	0.3237	0.3125	0.2060	0.2398
	CFSS [72]	0.3659	0.0632	0.3157	0.3854	0.3691	0.2688	0.3037
	DVLN [53]	0.4551	0.1474	0.3889	0.4743	0.4494	0.3794	0.3973
	LAB	**0.5323**	**0.2345**	**0.4951**	**0.5433**	**0.5394**	**0.4490**	**0.4630**

Table 3: Evaluation of LAB and several state-of-the-arts on Testset and 6 typical subsets of WFLW (98 landmarks).

Method	Mean Error (%)	Failure Rate (%)
Human	5.6	-
RCPR [6]	8.50	20.00
HPM [19]	7.50	13.00
CCR [17]	7.03	10.9
DRDA [68]	6.46	6.00
RAR [55]	6.03	4.14
SFPD [54]	6.40	-
DAC-CSR [18]	6.03	4.73
LAB w/o boundary	5.58	2.76
LAB	**3.92**	**0.39**

(a) Mean error (%) on COFW-29 testset.

Method	AFLW-Full (%)	AFLW-Frontal (%)
CDM [64]	5.43	3.77
RCPR [6]	3.73	2.87
ERT [26]	4.35	4.35
LBF [38]	4.25	2.74
CFSS [72]	3.92	2.68
CCL [73]	2.72	2.17
TSR [31]	2.17	-
DAC-OSR [18]	2.27	1.81
LAB w/o boundary	1.85	1.62
LAB	**1.25**	**1.14**

(b) Mean error (%) on AFLW testset.

Table 4: Cross-dataset evaluation on COFW and AFLW.

error on the Challenging subset which reflects the effectiveness of handling large head rotation and exaggerated expressions. Apart from 300W Fullset, we also show our results on 300W Testset in Table 2. Our method performs best among all of the state-of-the-art methods.

To verify the effectiveness and potential of boundary maps, we use ground truth boundary in the proposed method and report results named "LAB+oracle" which significantly outperform all the methods. The results demonstrate the effectiveness of boundary information and show great potential performance gain if the boundary information can be well captured.

4.1.2 Evaluation on WFLW

For comprehensively evaluating the robustness of our method, we report mean error, failure rate and AUC on the Testset and six typical subsets of WFLW on Table. 3. These six subsets were split from Testset by the provided attribute annotations. Though reasonable performance is obtained, there is illustrated to be still a lot of room for improvement for the extreme diversity of samples on WFLW, *e.g.*, large pose, exaggerated expressions and heavy occlusion.

4.1.3 Cross-dataset evaluation on COFW and AFLW

COFW-68 is produced by re-annotating COFW dataset with 68 landmarks annotation scheme to perform cross-dataset experiments by [19]. Fig. 7 shows the CED curves of our method against state-of-the-art methods on the COFW-68 [19] dataset. Our model outperforms previous results
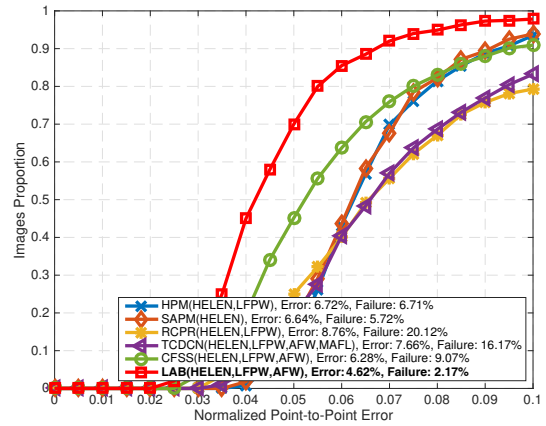

Figure 7: CED for COFW-68 testset (68 landmarks). Train set (in parentheses), mean error and failure rate are also reported.

with a large margin. We achieve 4.62% mean error with 2.17% failure rate. The failure rate is significantly reduced by 3.75%, which indicates the robustness of our method to handle occlusions.

In order to verify the capacity of handling cross-dataset face alignment of our method, we use boundary heatmaps estimator trained on 300W Fullset which has no overlap with COFW and AFLW dataset and compare the performance with and without using boundary information fusion ("LAB w/o boundary"). The results are reported in Table 4. The performance of previous methods without using 300-W datasets is also attached as a reference. There is a clear boost between our method without and with using boundary information. Thanks to the generalization of facial boundaries, the estimator learned on 300W can be conveniently used to supply boundary information for coordinate regression on COFW-29 [5] and AFLW [28] dataset, even though these datasets have different annotation protocols.

Moreover, our method uses boundary information achieves 29%, 32% and 29% relative performance improve-

ment over the baseline method ("LAB without boundary") on COFW-29, AFLW-Full and AFLW-Frontal respectively. Since COFW covers different level of occlusion and AFLW has significant view changes and challenging shape variations, the results emphasise the robustness brought by boundary information to occlusion, pose and shape variations. More qualitative results are demonstrated in our supplementary material.

4.2. Ablation study

Our framework consists of several pivotal components, *i.e.*, boundary information fusion, message passing and adversarial learning. In this section, we validate their effectiveness within our framework on the 300W Challenging Set and WFLW Dataset. Based on the baseline res-18 network (BL), we analyse each proposed component, *i.e.*, with the baseline hourglass boundary estimator ("HBL"), message passing ("MP"), and adversarial learning ("AL"), by comparing their mean error and failure rate. The overall results are shown in Fig. 8.

Boundary information is chosen as geometric structure representation in our work. We verify the potential of other structure information as well, *i.e.*, facial parts gaussian ("FPG") and face parsing results ("FP"). We report the landmarks accuracy with oracle results in Table 5 using different structure information. It can be observed easily that boundary map ("BM") is the most effective one.

Boundary information fusion is one of the key steps in our algorithm. We can fuse boundary information at different levels for the regression network. As indicated in Table 6, our final model that fuses boundary information in all four levels improves mean error from 7.12% to 6.13%. To evaluate the relationship between the quantity of boundary information fusion and the final prediction accuracy, we vary the number of fusion levels from 1 to 4 and report the mean error results in Table 6. It can be observed that performance is improved consistently by fusing boundary heatmaps at more levels.

Method	BL	BL+FPG	BL+FP	BL+BM
Mean Error	7.12	5.25	4.16	**3.28**

Table 5: Mean error (%) on 300W Challenging Set for evaluation the potential of boundary map as the facial structure information.

Method	BL	BL+L1	BL+L1&2	BL+L1&2&3	BL+L1&2&3&4
Mean Error	7.12	6.56	6.32	6.19	**6.13**

Table 6: Mean error (%) on 300W Challenging Subset for various fusion levels.

Method	BL	BL+HG/B	BL+CL	BL+HG
Mean Error	7.12	6.95	6.24	**6.13**

Table 7: Mean error (%) on 300W Challenging Set for different settings of boundary fusion scheme.

Method	HBL	HBL+MP	HBL+MP+AL
Error of heatmap	0.85	0.76	**0.63**
Error of landmark	6.13	5.82	**5.59**

Table 8: Normalised pixel-to-pixel error (%) of heatmap estimation, mean error (%) and failure rate (%) of landmark prediction on 300W Challenging Set for evaluation the relationship between the quality of estimated boundary and final prediction.
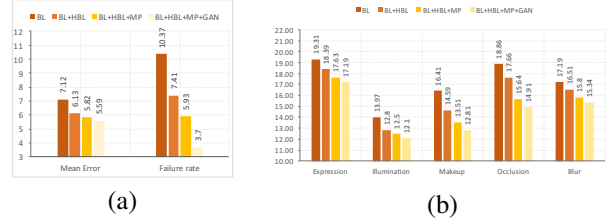

(a)   (b)

Figure 8: (a) Mean error (%) and failure rate (%) on 300W Challenging Subset. (b) Mean error (%) on 5 typical testing subset of WFLW Dataset, *i.e.* Expression, Illumination, Makeup, Occlusion and Blur Subset.

To verify the effectiveness of the fusion scheme shown in Fig. 4, we report the results of mean error on several settings in Table 7, *i.e.*, the baseline res-18 network ("BL"), hourglass module without boundary feature ("HG/B"), hourglass module with boundary feature ("HG") and consecutive convolutional layers with boundary feature ("CL"). The comparison between "BL+HG" and "BL+HG/B" indicates the effectiveness of boundary information fusion rather than network structure changes. The comparison between "BL+HG" and "BL+CL" indicates the effectiveness of the using hourglass structure design.

Message passing plays a vital role for heatmap quality improvement when severe occlusions happen. As illustrated in Fig. 8 (b) on Occlusion Subset of WFLW, message passing, which combines information from visible boundaries and occluded ones, reduce the mean error over 11% relatively.

Adversarial learning further improves the quality and effectiveness of boundary heatmaps. As illustrated in Fig. 5, heatmaps can be observed to be more focused and salience when adversarial loss is added. To verify the effectiveness of our landmark based boundary effectiveness discriminator, a baseline method using traditionally defined discriminator is tested on 300W Challenging Set. The failure rate is reduced from 5.19% to 3.70%.

Relationship between boundary estimator and landmarks regressor is evaluated by analyzing the quality of estimated heatmap and final prediction accuracy. We report the MSE of estimated heatmaps and corresponding landmarks accuracy in Table 8. We observe that with message passing ("HBL+MP") and adversarial learning ("HBL+AL"), the errors of estimated heatmaps are reduced together with landmarks accuracy.

5. Conculsion

Unconstrained face alignment is an emerging topic. In this paper, we present a novel use of facial boundary to derive facial landmarks. We believe the reasoning of a unique facial structure is the key to localise facial landmarks, since human face does not include ambiguities. By estimating facial boundary, our method is capable of handling arbitrary head poses as well as large shape, appearance, and occlusion variations. Our experiment shows the great potential of modeling facial boundary. The runtime of our algorithm is $60ms$ on TITAN X GPU.

References

[1] R. Alp Guler, G. Trigeorgis, E. Antonakos, P. Snape, S. Zafeiriou, and I. Kokkinos. Densereg: Fully convolutional dense shape regression in-the-wild. In *CVPR*, 2017.

[2] K. Amit and C. Rama. Disentangling 3d pose in a dendritic cnn for unconstrained 2d face alignment. *arXiv preprint*, arXiv:1802.06713, 2018.

[3] P. N. Belhumeur, D. W. Jacobs, D. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *CVPR*, 2011.

[4] A. Bulat and Y. Tzimiropoulos. Convolutional aggregation of local evidence for large pose face alignment. In *BMVC*, 2016.

[5] X. P. Burgos-Artizzu, P. Perona, and P. Dollar. Robust face landmark estimation under occlusion. In *ICCV*, 2013.

[6] X. P. Burgos-Artizzu, P. Perona, and P. Dollár. Robust face landmark estimation under occlusion. In *ICCV*, 2013.

[7] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. *IJCV*, 107(2):177–190, 2014.

[8] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun. Joint cascade face detection and alignment. In *ECCV*, 2014.

[9] Y. Chen, C. Shen, X.-S. Wei, L. Liu, and J. Yang. Adversarial posenet: A structure-aware convolutional network for human pose estimation. In *ICCV*, 2017.

[10] C. Chou, J. Chien, and H. Chen. Self adversarial training for human pose estimation. *arXiv preprint*, arXiv:1707.02439, 2017.

[11] X. Chu, W. Ouyang, H. Li, and X. Wang. Structured feature learning for pose estimation. In *CVPR*, 2016.

[12] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang. Multi-context attention for human pose estimation. In *CVPR*, 2017.

[13] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(6):681–685, 2001.

[14] J. Deng, Q. Liu, J. Yang, and D. Tao. M^3 CSR: multi-view, multi-scale and multi-component cascade shape regression. *Image Vision Comput.*, 47:19–26, 2016.

[15] J. Deng, G. Trigeorgis, Y. Zhou, and S. Zafeiriou. Joint multi-view face alignment in the wild. *arXiv preprint*, arXiv:1708.06023, 2017.

[16] H. Fan and E. Zhou. Approaching human level facial landmark localization by deep learning. *Image Vision Comput.*, 47:27–35, 2016.

[17] Z. Feng, G. Hu, J. Kittler, W. J. Christmas, and X. Wu. Cascaded collaborative regression for robust facial landmark detection trained using a mixture of synthetic and real images with dynamic weighting. *TIP*, 24(11):3425–3440, 2015.

[18] Z.-H. Feng, J. Kittler, W. Christmas, P. Huber, and X.-J. Wu. Dynamic attention-controlled cascaded shape regression exploiting training data augmentation and fuzzy-set sample weighting. In *CVPR*, 2017.

[19] G. Ghiasi and C. C. Fowlkes. Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model. In *CVPR*, 2014.

[20] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial networks. *arXiv preprint*, arXiv:1406.2661, 2014.

[21] T. Hassner, S. Harel, E. Paz, and R. Enbar. Effective face frontalization in unconstrained images. In *CVPR*, 2015.

[22] Z. Hayder, X. He, and M. Salzmann. Boundary-aware instance segmentation. In *CVPR*, 2017.

[23] D. C. Hogg and R. Boyle, editors. *Proceedings of the British Machine Vision Conference, BMVC 1992, Leeds, UK, September, 1992*. BMVA Press, 1992.

[24] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint*, arXiv:1408.5093, 2014.

[25] F. Kahraman, M. Gökmen, S. Darkner, and R. Larsen. An active illumination and appearance (AIA) model for face alignment. In *CVPR*, 2007.

[26] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *CVPR*, 2014.

[27] A. Kirillov, E. Levinkov, B. Andres, B. Savchynskyy, and C. Rother. Instancecut: From edges to instances with multicut. In *CVPR*, 2017.

[28] M. Köstinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *ICCV Workshop*, 2011.

[29] N. Kumar, P. Belhumeur, and S. Nayar. Facetracer: A search engine for large collections of images with faces. In *ECCV*, 2008.

[30] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *ECCV*, 2012.

[31] J. Lv, X. Shao, J. Xing, C. Cheng, and X. Zhou. A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In *CVPR*, 2017.

[32] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla. Classification with an edge: Improving semantic image segmentation with boundary detection. *arXiv preprint*, arXiv:1612.01337, 2016.

[33] I. A. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60(2):135–164, 2004.

[34] S. Milborrow and F. Nicolls. Locating facial features with an extended active shape model. In *European Conference on Computer Vision*, pages 504–513, 2008.

[35] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016.

[36] A. Opelt, A. Pinz, and A. Zisserman. A boundary-fragment-model for object detection. In *ECCV*, 2006.

[37] H. Pan and L. Xia. Comic: Good features for detection and matching at object boundaries. *arXiv preprint*, arXiv:1412.1957, 2014.

[38] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 FPS via regressing local binary features. In *CVPR*, 2014.

[39] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *ICCV Workshop*, 2013.

[40] M. S. M. Sajjadi, B. Scholkopf, and M. Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *ICCV*, 2017.

[41] J. M. Saragih and R. Göcke. A nonlinear discriminative approach to AAM fitting. In *ICCV*, 2007.

[42] J. M. Saragih, S. Lucey, and J. F. Cohn. Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision*, 91(2):200–215, 2011.

[43] W. Shen, B. Wang, Y. Jiang, Y. Wang, and A. Yuille. Multi-stage multi-recursive-input fully convolutional networks for neuronal boundary detection. In *ICCV*, 2017.

[44] W. Shen, X. Wang, Y. Wang, X. Bai, and Z. Zhang. Deepcontour: A deep convolutional feature learned by positive-sharing loss for contour detection. In *CVPR*, 2015.

[45] C. C. L. Shuo Yang, Ping Luo and X. Tang. From facial parts responses to face detection: A deep learning approach. In *ICCV*, 2015.

[46] B. M. Smith and L. Zhang. Collaborative facial landmark localization for transferring annotations across datasets. In *ECCV*, 2014.

[47] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *CVPR*, 2013.

[48] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10, 000 classes. In *CVPR*, 2014.

[49] Y. Sun, X. Wang, and X. Tang. Hybrid deep learning for face verification. *TPAMI*, 38(10):1997–2009, 2016.

[50] A. Toshev, B. Taskar, and K. Daniilidis. Shape-based object detection via boundary structure segmentation. *IJCV*, 99(2):123–146, 2012.

[51] G. Trigeorgis, P. Snape, M. A. Nicolaou, E. Antonakos, and S. Zafeiriou. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *CVPR*, 2016.

[52] J. R. R. Uijlings and V. Ferrari. Situational object boundary detection. In *CVPR*, 2015.

[53] W. Wu and S. Yang. Leveraging intra and inter-dataset variations for robust face alignment. In *CVPR Workshop*, 2017.

[54] Y. Wu, C. Gou, and Q. Ji. Simultaneous facial landmark detection, pose and deformation estimation under facial occlusion. In *CVPR*, 2017.

[55] S. Xiao, J. Feng, J. Xing, H. Lai, S. Yan, and A. A. Kassim. Robust facial landmark detection via recurrent attentive-refinement networks. In *ECCV*, 2016.

[56] S. Xie and Z. Tu. Holistically-nested edge detection. In *ICCV*, 2015.

[57] X. Xiong and F. D. la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, 2013.

[58] X. Xiong and F. D. la Torre. Global supervised descent method. In *CVPR*, 2015.

[59] D. Xuanyi, Y. Yan, O. Wanli, and Y. Yi. Style aggregated network for facial landmark detection. *arXiv preprint*, arXiv:1803.04108, 2018.

[60] J. Yang, Q. Liu, and K. Zhang. Stacked hourglass network for robust facial landmark localisation. In *CVPR Workshop*, 2017.

[61] S. Yang, P. Luo, C. C. Loy, and X. Tang. Wider face: A face detection benchmark. In *CVPR*, 2016.

[62] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang. Learning feature pyramids for human pose estimation. In *ICCV*, 2017.

[63] W. Yang, W. Ouyang, H. Li, and X. Wang. End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. In *CVPR*, 2016.

[64] X. Yu, J. Huang, S. Zhang, W. Yan, and D. N. Metaxas. Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In *ICCV*, 2013.

[65] Z. Yu, C. Feng, M.-Y. Liu, and S. Ramalingam. Casenet: Deep category-aware semantic edge detection. In *CVPR*, 2017.

[66] S. Zafeiriou, G. Trigeorgis, G. Chrysos, J. Deng, and J. Shen. The menpo facial landmark localisation challenge: A step towards the solution. In *CVPR Workshop*, 2017.

[67] J. Zhang, M. Kan, S. Shan, and X. Chen. Leveraging datasets with varying annotations for face alignment via deep regression network. In *ICCV*, 2015.

[68] J. Zhang, M. Kan, S. Shan, and X. Chen. Occlusion-free face alignment: Deep regression networks coupled with decorrupt autoencoders. In *CVPR*, 2016.

[69] J. Zhang, S. Shan, M. Kan, and X. Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *ECCV*, 2014.

[70] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Learning deep representation for face alignment with auxiliary attributes. *TPAMI*, 38(5):918–930, 2016.

[71] S. Zhu, C. Li, C. C. Loy, and X. Tang. Transferring landmark annotations for cross-dataset face alignment. *CoRR*, abs/1409.0602, 2014.

[72] S. Zhu, C. Li, C. C. Loy, and X. Tang. Face alignment by coarse-to-fine shape searching. In *CVPR*, 2015.

[73] S. Zhu, C. Li, C. C. Loy, and X. Tang. Unconstrained face alignment via cascaded compositional learning. In *CVPR*, 2016.

[74] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. In *CVPR*, 2016.

[75] Z. Zhu, P. Luo, X. Wang, and X. Tang. Deep learning identity-preserving face space. In *ICCV*, 2013.