Look at Boundary: A Boundary-Aware Face Alignment Algorithm
Supplementary Material

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

In this document, to facilitate future reimplementation of our work, more details of experiment and message passing scheme are provided. Additional qualitative results, which are not covered in the submission due to the page limit, are reported. Moreover, detail annotations and representative examples of the proposed WFLW Dataset are demonstrated.

1. Implementation Details

All training images are cropped and resized to 256×256 according to bounding boxes provided by datasets. Then, standard data augmentation is performed including translation (± 30 pixels), rotation (± 20 degrees), scaling ($\pm 15\%$) and flip to make the model more robust to data variations.

In this work, for comparison with state-of-the-arts, 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. To get our baseline regressor and effectiveness discriminator, a res-18 [2] network is modified by adding two fully-connected layers with 256 units for the final prediction. All our models are trained with *Caffe*. We use stochastic gradient descent (SGD) to optimise the network on 4 Titan X GPUs with a mini-batch size of 8 for 2000 epochs. We set weight decay and momentum equal to 0.0005 and 0.9 respectively. The learning rate is initialised as 2×10^{-5} and is dropped by 5 at the $1000th$ and the $1500th$ epoch. Empirically, σ is set to 1, θ is set to 1.5 and δ is set to 0.8 in our experiments. Details of all evaluation settings for different experiments are noted in Table 1.

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

2. Details of Message Passing

Specifically, the message passing layers can be effectively implemented by a series of convolution and entry-wise sum operation. With an attached loss in each stack of hourglass, these message layers can be jointly learnt with kernels of the network. As illustrated in Fig. 1, denote $\mathbf{h}_t$ as the feature with 256 channels obtained by a 1×1 convolution at the end of stack t. Each boundary i in stack t before message passing can be represented as feature $\mathbf{A}_i^t$ with 16 channels calculated as follows:

$$\mathbf{A}_i^t = f(\mathbf{h}_t * \mathbf{w}^{a_i^t}) \tag{1}$$

where $\mathbf{w}^{a_i^t}$ denotes the filter bank for boundary i, $*$ denotes convolution and f is the nonlinear activation function. The refined feature for boundary i in stack t after message passing is denoted by $\mathbf{A'}_i^t$.

As a concrete example, we demonstrate the positive message passing process to boundary 6 (left eyebrow) in stack 2. Following the tree structure of our method, left eyebrow in stack 2 catches messages from two nodes, *i.e.*, boundary 2 (left upper eyelid) in stack 2 and boundary 6 (left eyebrow) in stack 1. Thus, the refined feature for boundary 6 in stack 2 after receiving information from intra and inter-level boundaries can be formulated as follows:

$$\mathbf{A'}_6^2 = f(\mathbf{A}_6^2 + \mathbf{A'}_2^2 * \mathbf{w}^{a_2^2, a_6^2} + \mathbf{A'}_6^1 * \mathbf{w}^{a_6^1, a_6^2}) \tag{2}$$

where $+$ donetes entrywise sum, $\mathbf{A}_6^2$, which can be calculated by Eq. 1, denotes the feature before updated. $\mathbf{A'}_2^2 * \mathbf{w}^{a_2^2, a_6^2}$ denotes the meassage comes from intra-level boundary $\mathbf{A'}_2^2$, by a collection of convolution kernels $\mathbf{w}^{a_2^2, a_6^2}$. $\mathbf{A'}_6^1 * \mathbf{w}^{a_6^1, a_6^2}$ denotes the meassage comes from inter-level boundary $\mathbf{A'}_6^1$, by $\mathbf{w}^{a_6^1, a_6^2}$. Also, $\mathbf{A'}_2^2$ and $\mathbf{A'}_6^1$ can be iteratively calculated until meeting the leaf node of the tree model.

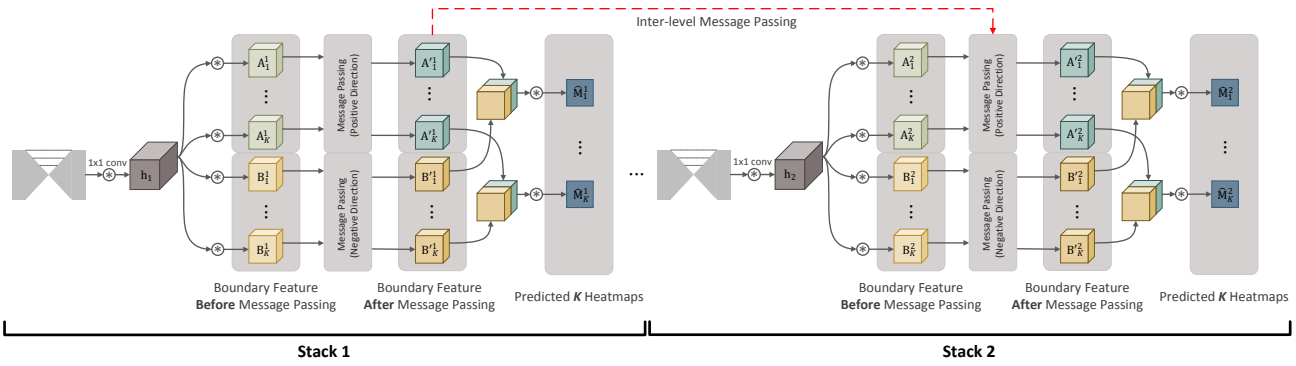

Figure 1: An illustration of the detail of message passing scheme. We take stack 1 and stack 2 for example.

A same tree structure with opposite direction is also used to pass information in a reverse flow. The refined feature $\mathbf{B'}_i^t$ obtained by opposite direction tree and $\mathbf{A'}_i^t$ are concatenated together to get the final predicted heatmap as follows:

$$\hat{M}_i^t = (\mathbf{A'}_i^t \oplus \mathbf{B'}_i^t) * \mathbf{w'}^{a_i^t} \tag{3}$$

where $\oplus$ represents concatenation, $\mathbf{w'}^{a_i^t}$ is a 1×1 convolution filter bank to get the single channel predicted heatmap for boundary i in stack t. Each stack of hourglass network predicts K heatmaps $\hat{M}^t = \{\hat{M}_i\}_{i=1}^K$ and a loss is attached at the end of each stack defined by the Mean Squared Error (MSE):

$$\mathcal{L}_G = \frac{1}{2N} \sum_{i=1}^{N} \sum_{i=1}^{K} \|\hat{M}_i^t - M_i^t\|_2^2 \tag{4}$$

where N is the number of samples.

3. Qualitative Evaluation of Boundary Cues

To demonstrate the guidance effect of boundary information on the learning of landmarks regressor, we visualise the feature maps sampled at two stages with resolution 64×64 and 32×32 respectively. For comparison, visualisation results of a "Res-18" baseline network without boundary information are also shown in Fig. 2. It can be clear seen that, with boundary information incorporated in, feature maps tend to be more focused on facial boundaries than the baseline network. These boundary-guided feature maps will greatly ease the learning of landmarks regressor and thus enhance the efficiency of network parameters.

4. Qualitative Results

Additional qualitative results are illustrated in this section. As shown in Fig. 3, even with severe occlusion, the estimated boundary heatmaps of our method are still very plausible and focused. With the help of boundary cues, the predicted landmarks are also shown to be robust to the occluded parts. To evaluation of the robustness of our method to challenging variations of expression, illumination and view changes, we demonstrate the estimated boundary heatmaps and predicted landmarks on 300W Challenging Set in Fig. 4. Both of the boundary and landmark results are shown to be robust to several representative hard examples. Moreover, visualisation results of landmark localisation on COFW-29 (29 landmarks) and AFLW-Full (19 landmarks) with the help of cross-dataset boundary cues are shown in Fig. 5 and Fig. 6 respectively.

5. Wider Facial Landmark in the Wild

On the purpose of facilitating future research of face alignment, we introduce a new facial dataset base on WIDER Face [5] named Wider Facial Landmarks in the Wild (WFLW), which contains 10,000 faces with 98 fully manual annotated landmarks. The multi-view presentation and location of the proposed 98 points annotation are shown in Fig. 7. Also, WFLW provides rich property annotations, including pose, expression, illumination, make-up, occlusion and blur for comprehensive analysis of existing algorithms. As shown in Fig. 8, faces in the proposed dataset are all collected under unconstrained conditions and extremely challenging due to large variations in expression, pose and occlusion. We can simply evaluate robustness of pose, occlusion, and expression on proposed dataset instead of switching between multiple evaluation protocols in different datasets. The comparison of WFLW with the most widely used in-the-wild benchmark is shown in Table 2.

References

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

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Evaluation Name	Training Set	# of Training Samples	Testing Set	# of Testing Samples	Point	Normalising Factor
300W Fullset	300W train-set	3,148	300W full-set	689	68	inter-ocular/pupil distance
300W Testset	300W train-set	3,148	300W test-set	600	68	inter-ocular distance
COFW-68	300W train-set	3,148	COFW test-set	507	68	inter-ocular distance
COFW-29	COFW train-set	1,345	COFW test-set	507	29	inter-ocular distance
AFLW-Full	AFLW train-set	20,000	AFLW test-set	4,386	19	face size
AFLW-Frontal	AFLW train-set	20,000	AFLW frontal-set	1,314	19	face size
WFLW	WFLW train-set	7,500	WFLW test-set	2,500	98	inter-ocular distance

Table 1: Experiments Setting

Dataset	# Training	# Testing	# Landmarks	Pose	Expression	Illumination	Make-Up	Occlusion	Blur
AFLW [3]	20,000	4,386	21	-	-	-	-	-	-
300W [4]	3,148	1,289	68	-	-	-	-	-	-
COFW [1]	1,345	507	29	-	-	-	-	✓	-
WFLW	7,500	2,500	98	✓	✓	✓	✓	✓	✓

Table 2: Comparison of the most widely used face alignment datasets.

[3] 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.

[4] 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.

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

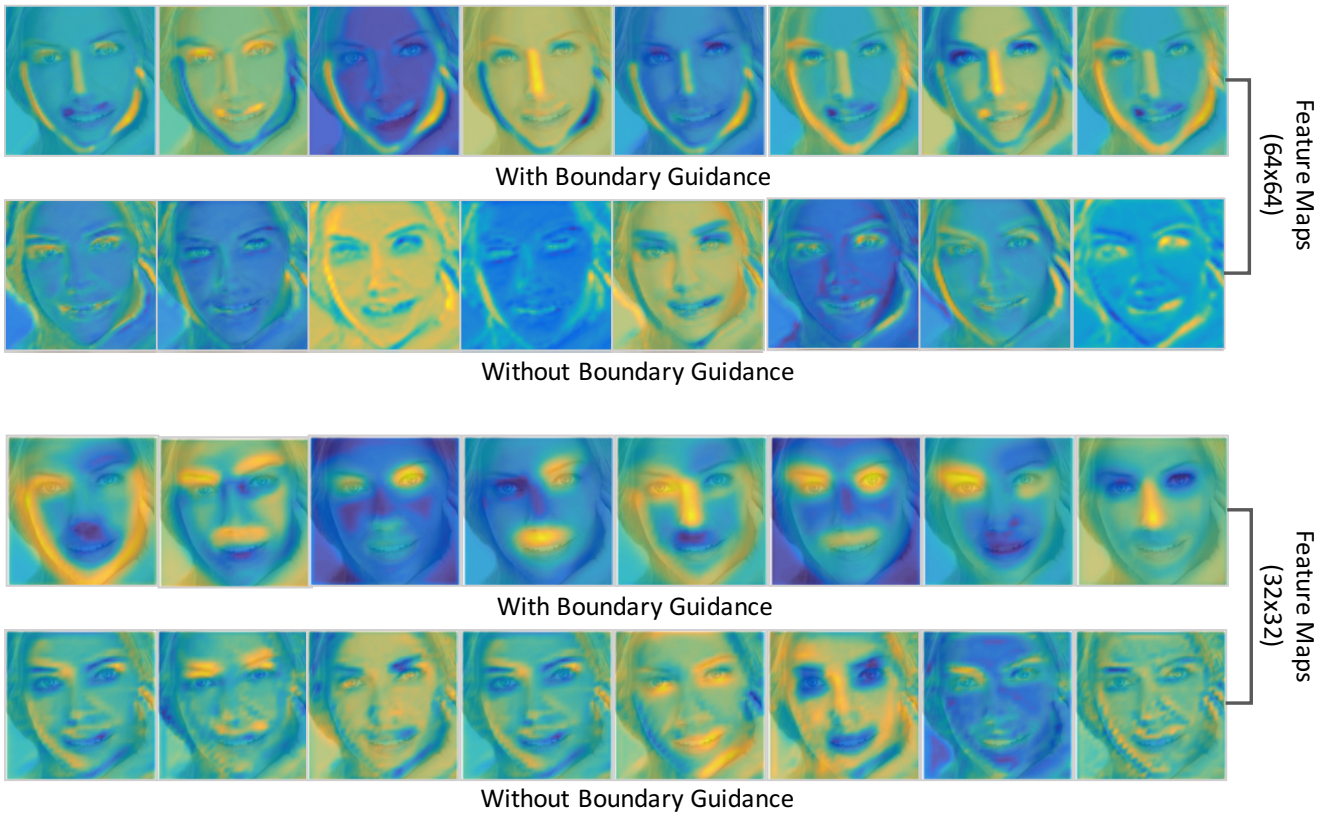

Figure 2: Feature maps with size 64×64 and 32×32 sampled randomly in the same layers of baseline res-18 and our proposed boundary-aware landmarks regressor. Feature maps are resized to the same size with face image and stcked together with it for illustration. (Best viewed in color)

Figure 3: Representative examples of boundary heatmap estimation and landmark localisation results on COFW-68 Testset. (Best viewed in color)

Figure 4: Representative examples of boundary heatmap estimation and landmark localisation results on 300W Challenging Set. (Best viewed in color)

Figure 5: Representative examples of landmark localisation results on COFW-29 Testset (29 landmarks).

Figure 6: Representative examples of landmark localisation results on AFLW-Full Testset (19 landmarks).

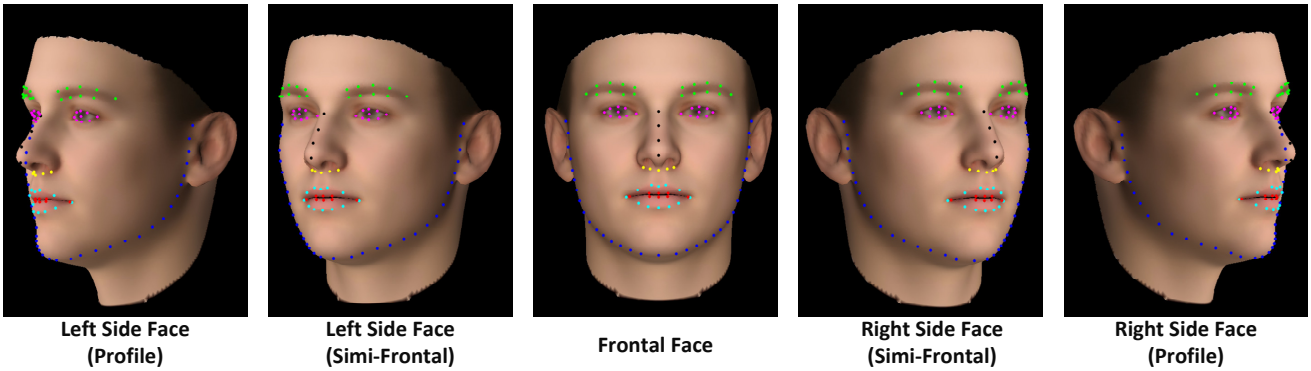

Figure 7: An illustration of the proposed 98 landmark annotations of WFLW Dataset in multi-view.

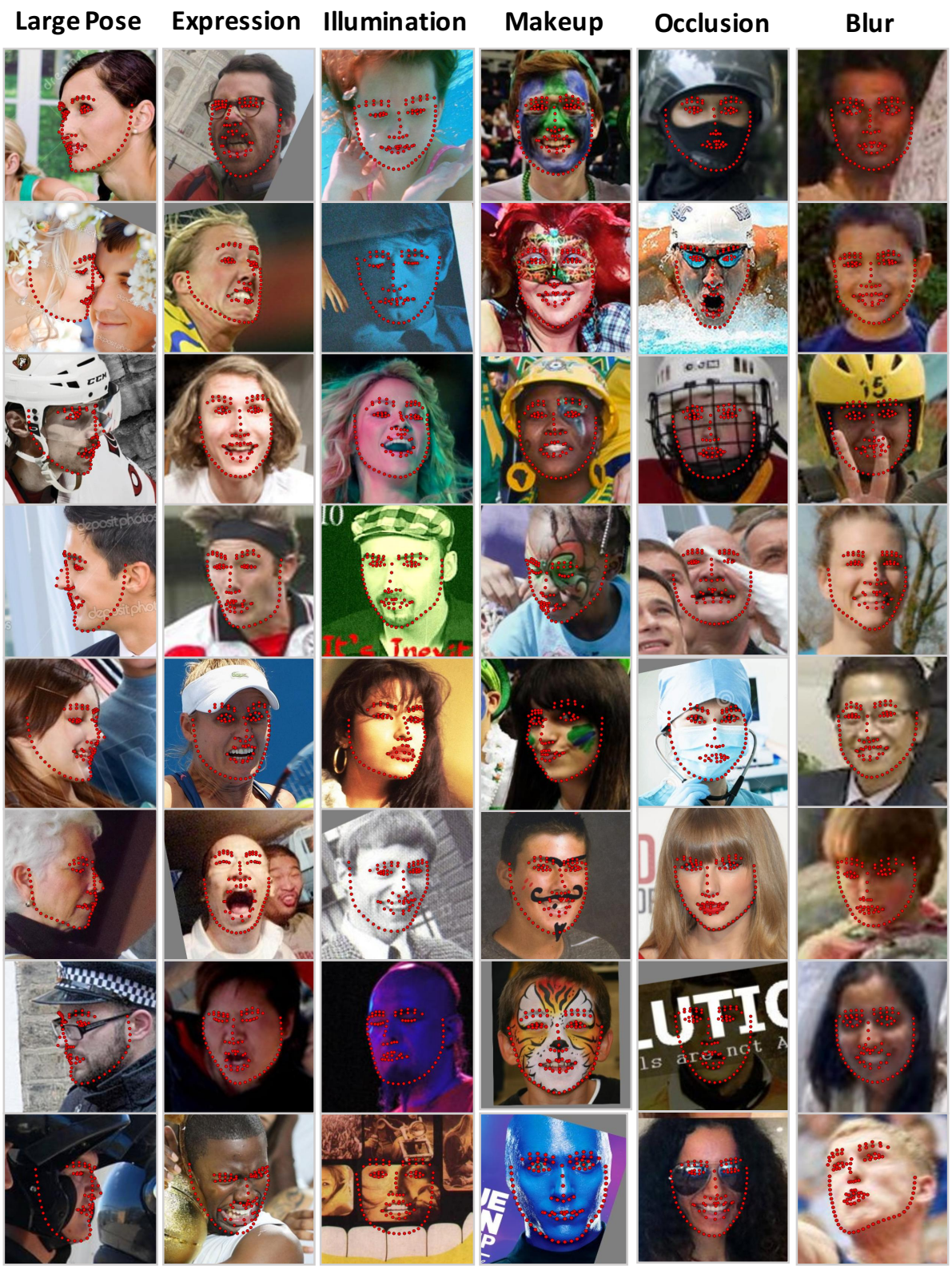

Figure 8: Representative examples of landmark annotations on 6 Testsets of WFLW Dataset.