Supplementary Materials

Yangyu Huang, Hao Yang[*], Chong Li, Jongyoo Kim, Fangyun Wei
Microsoft Research Asia
{yanghuan,haya,chol,jongk,fawe}@microsoft.com

In this supplementary material, we provide detailed architectures of the proposed ADNet, extended experimental results, comprehensive explanations about the anisotropic attention module. We also demonstrate how error-bias towards normal direction of face alignment leverages model training.

1. Model Architecture

Tables 1, 2 and 3 fully demonstrate the architecture of the proposed ADNet. For detailed introduction of our experimental setting, please refer to Section 4 of our manuscript. In the table, $P*$, $H*_{point}$ and $H*_{edge}$ denote the inputs of *smooth ADL1* loss, *AWing* loss and *AWing* loss, respectively. N_{point} and N_{edge} indicate the number of points and edges, which varies according to each dataset. The loss weights of Hour Glass (HG) for stacked 4 HGs are respectively 1/8, 1/4, 1/2, and 1. The fourth head branch outputs P_3 is the final predicted coordinate of each landmark, which is derived from the soft argmax operation.

In Table 2, the goal of $E2P$ *Transform* is to convert $\hat{H}_{edge}$ (N_{edge} channels) into H_{edge} (N_{point} channels) by considering the adjacency relationship as

$$E2P\ \textit{Transform}(\hat{H}_{edge}(x, y)) = \textit{Mat}_{E2P} \cdot \hat{H}_{edge}(x, y) \quad (1)$$

where $\hat{H}_{edge}(x, y)$ is a column vector at the position of (x, y), and $\textit{Mat}_{E2P}$ is a $N_{point} \times N_{edge}$ binary matrix describing the adjacency relationship between each point and each edge. More specifically, if the ith point is connected to the jth edge, $\textit{Mat}_{E2P}(i, j) = 1$, otherwise, $\textit{Mat}_{E2P}(i, j) = 0$. Note that $\textit{Mat}_{E2P}$ is a constant variable, and is derived based on the landmark definition of each database, respectively.

2. Edge Definition

We categorize the landmarks into two groups: *edge landmarks* and *point landmarks*. If the landmarks locate on edges, they belong to the former group, conversely, landmarks not on edges belong to the latter group. For several well-known face alignment datasets such as COFW, 300W, and WFLW, most of the landmarks belong to edge landmarks. We show our definition of edges in 300W dataset in Table 4 and Figure 1.

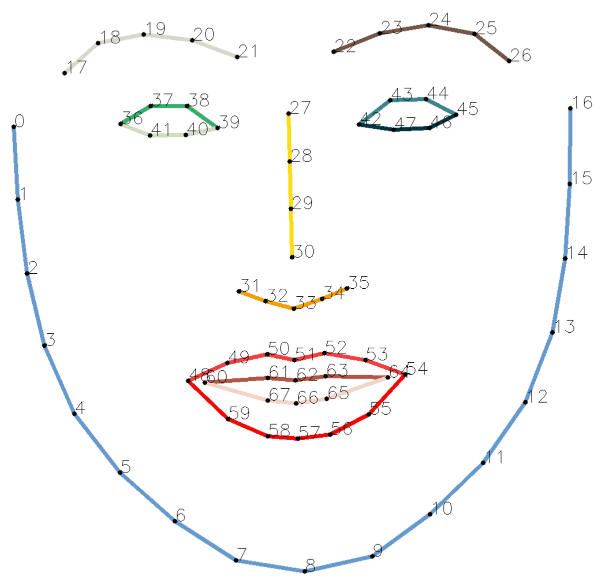

Figure 1. Visualized example of edges in 300W. Each colored line corresponds to each edge defined in Table 4.

3. Additional Experiments and Results

3.1. Comparison of Inference Time

To show the computational complexity of ADL and AAM, we compare the inference time of the baseline model and ADNet. Note that the baseline model is almost identical to ADNet except that AAM and ADL are removed from the baseline. To estimate the time, we repeated the experiment 10 times on the 300W fullset and averaged the measured times. We used one NVIDIA v100 GPU with a batch size of 1. As tabulated in Table 5, ADNet takes only 6% longer time than the baseline method, which indicates the high efficiency of ADL and AAM. Moreover, ADL and AAM take small FLOPs and require a small number of parameters as shown in the table.

[*]Corresponding author

Layer	Input of layer	Output of layer	Output Channels	Kernel Size	Stride	Padding
Input	image	-	-	-	-	-
Coord Conv [2]	image	x0	64	7	2	3
BN-ReLu	x0	x1	64	-	-	-
Residual Block [1]	x1	x2	128	-	-	-
Max Pool	x2	x3	128	2	2	0
Blur Pool [4]	x3	x4	128	3	2	0
Residual Block	x4	x5	128	-	-	-
Residual Block	x5	x6	256	-	-	-
Head Branch	x6	($P0$, x7, $H0_{point}$, $H0_{edge}$)	-	-	-	-
Head Branch	x7	($P1$, x8, $H1_{point}$, $H1_{edge}$)	-	-	-	-
Head Branch	x8	($P2$, x9, $H2_{point}$, $H2_{edge}$)	-	-	-	-
Head Branch	x9	($P3$, x10, $H3_{point}$, $H3_{edge}$)	-	-	-	-
Output	-	($P*$, $H*_{point}$, $H*_{edge}$)	-	-	-	-

Table 1. The architecture of ADNet. x[*] and H[*] indicate intermediate feature maps, and BN indicates batch normalization. The detailed structure of "Head Branch" and "Residual Block" are shown in Tables 2 and 3.

Layer	Input of layer	Output of layer	Output Channels	Kernel Size	Stride	Padding
Input	y0	-	-	-	-	-
Hour Glass [3]	y0	y1	256	-	-	-
Conv-BN-ReLu	y1	y2	256	1	1	0
Residual Block	y2	y3	256	-	-	-
Conv-Sigmoid	y3	H_{point}	N_{point}	1	1	0
Conv-Sigmoid	y3	$\hat{H}_{edge}$	N_{edge}	1	1	0
E2P Transform	$\hat{H}_{edge}$	H_{edge}	N_{point}	-	-	-
Elementwise dot	(H_{point}, H_{edge})	$H_{point-edge}$	N_{point}	-	-	-
Conv-ReLu	y3	$H_{landmarks}$	N_{point}	1	1	0
Elementwise dot	($H_{landmarks}$, $H_{point-edge}$)	$AH_{landmarks}$	N_{point}	-	-	-
Soft Argmax	$AH_{landmarks}$	P	N_{point}	-	-	-
Conv	$H_{landmarks}$	y4	256	1	1	0
Conv	H_{point}	y5	256	1	1	0
Conv	H_{edge}	y6	256	1	1	0
Elementwise sum	(y3, y4, y5, y6)	y7	256	-	-	-
Output	-	(P, y7, H_{point}, H_{edge})	-	-	-	-

Table 2. The architecture of head branch.

Layer	Input of layer	Output of layer	Output Channels	Kernel Size	Stride	Padding
Input	z0	-	-	-	-	-
BN-ReLu-Conv	z0	z1	output channels / 2	1	1	0
BN-ReLu-Conv	z1	z2	output channels / 2	3	1	1
BN-ReLu-Conv	z2	z3	output channels	1	1	0
Skip	z0	z4	output channels	1	1	0
Elementwise sum	(z3, z4)	z5	output channels	1	1	0
Output	-	z5	-	-	-	-

Table 3. The architecture of residual block. "output channels" denotes the channel size of the residual block's output.

Components	Edge Names	Vertex Indices
Contour	Face Contour	0-16
Eyebrow	Right Eyebrow	17-21
	Left Eyebrow	22-26
Nose	Nose Middle Line	27-30
	Nose Bottom Line	31-35
Eye	Right Eye Superior Margin	36-39
	Right Eye Inferior Margin	39-41, 36
	Left Eye Superior Margin	42-45
	Left Eye Inferior Margin	45-47, 42
Mouth	Outer Lip Superior Margin	48-54
	Outer Lip Inferior Margin	54-59, 48
	Inner Lip Superior Margin	60-64
	Inner Lip Inferior Margin	64-67, 60
Whole face	-	0-67

Table 4. Definition of edges in 300W. The visualized example of each edge is shown in Figure 1 with the same color.

Methods	Inference Time	FLOPs	Params
Baseline	89.49 ms/face	16.46G	13.23M
ADNet	95.29 ms/face	17.04G	13.37M

Table 5. The comparison of inference time, FLOPs and the number of parameters on the 300W fullset.

3.2. Evaluation of Individual Edges on 300W

Apart from evaluating the whole face on the test dataset, we also provide the NME of each edge in the 300W fullset dataset to fully demonstrate the effectiveness of the proposed method. The detailed results are shown in Table 7. The bias rate is defined as

$$Bias\ Rate = \frac{NME_{tangent} - NME_{normal}}{NME_{normal}} \quad (2)$$

where $NME_{tangent}$ and NME_{normal} are respectively the NME in tangent and normal directions. For both normal NME and tangent NME, ADNet outperforms the baseline method for every edge. In addition, ADNet has always larger bias rate than the baseline, which means that ADNet is leveraging the bias towards normal direction.

3.3. Exploration of λ Settings

We investigate three λ settings in Table 6: **i)** All landmarks have the same value $\lambda_i = 2$: (c)(f). Other λ_i can be found in Table 4 of our paper. **ii)** $\lambda_i = 4$ for the outer face contour (denoted by $\mathcal{O}$ in Table 6), and $\lambda_i = 2$ for the rest: (d)(g). **iii)** Independent λ_i for each landmark: (e)(h). Each was computed by $\lambda_i = a_i/b_i$, where a_i and b_i are long and short radius of each fitted ellipse by error distribution in Fig 1(a) of our paper.

It can be observed that: **i)** though a more flexible λ_i leads to better performance, the improvement is marginal; **ii)** the significant improvement comes from AAM rather than ADL.

width=0.8

ID	Components	λ_i	NME (%)
(a)	*Baseline*	-	3.38
(b)	AAM only	-	2.98
(c)	ADL only	$\lambda_i = 2$	3.231951
(d)	ADL only	$\lambda_{i\in\mathcal{O}} = 4, \lambda_{i\notin\mathcal{O}} = 2$	3.229207
(e)	ADL only	$\lambda_i = a_i/b_i$	3.219207
(f)	AAM + ADL	$\lambda_i = 2$	2.934116
(g)	AAM + ADL	$\lambda_{i\in\mathcal{O}} = 4, \lambda_{i\notin\mathcal{O}} = 2$	2.934933
(h)	AAM + ADL	$\lambda_i = a_i/b_i$	2.930612

Table 6. Evaluating different λ strategies on 300W in terms of interocular NME. The *Baseline* in (a) removes both AAM and ADL.

3.4. Demonstration of Error Distribution on 300W

To demonstrate the error-bias in error distribution with real-world data, in Figure 2, we provide the empirical error distribution of chin point obtained by using an off-the-shelf face alignment algorithm on the 300W dataset trained by baseline method. It is obvious that the error distribution along tangent direction (tangent distribution in figure) is broader than that along the normal direction (normal distribution in figure), which is consistent with our assumption, error-bias towards normal direction.

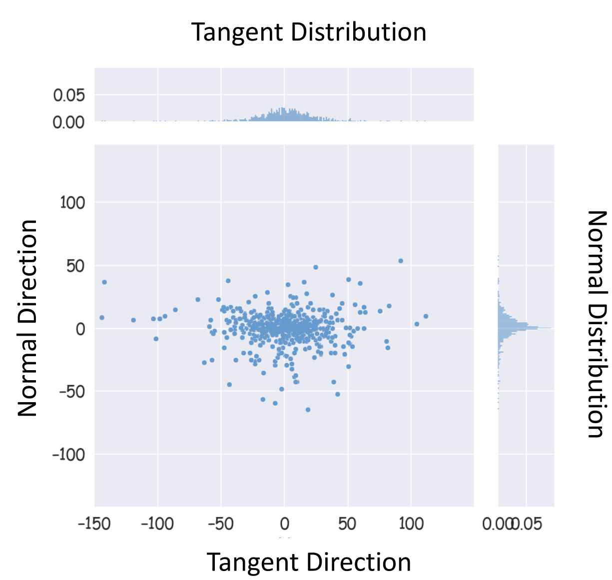

Figure 2. Error distribution of chin landmark (the $8th$ point in Figures 1) on the 300W fullset dataset obtained by off-the-shelf face alignment model. Each sub-figure (up/right) shows the projected error distribution along (tangent/normal) direction.

3.5. Visualized Examples of ADNet

To verify the robustness of ADNet, we additionally show the landmark inference on the extended test data in Figure 4, 5 and 6. For each image, the first row (red landmarks) is the inference result by ADNet and the second row (green landmarks) is the corresponding ground-truth provided by the dataset. As can be seen, our method yields stable and reasonable prediction of landmarks even for difficult cases such as extreme occlusion, large pose, extreme expression, blur and bad illumination.

4. Relationship between AAM and Proposed Guideline

As described in the manuscript, the anisotropic attention module outputs an anisotropic mask per landmarks. By design, the anisotropic mask has a strong response in tangent direction and a weak response in normal direction. Consequently, each predicted landmark has a large tolerance for tangent error, but small tolerance for normal error. This can be confirmed in the visualized example in Figure 3, where the AAM mask has broad distribution along tangent direction (ranging between t_0 to t_1) while the distribution along normal direction is limited (ranging between n_0 to n_1). In other words, the guideline imposes strong constraints along the normal direction of each landmark.

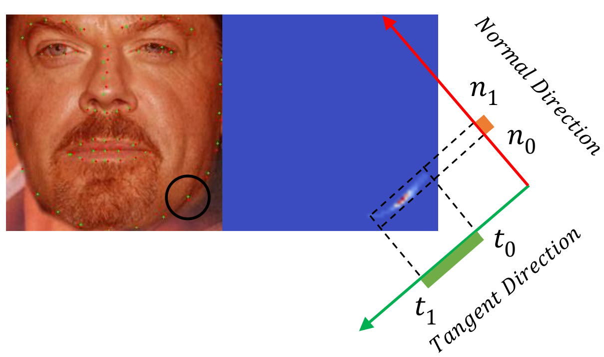

Figure 3. Error tolerance in different direction by applying AAM mask. The **orange** segment indicates the predicted coordinate range in normal direction, and **green** segment indicates the predicted coordinate range in tangent direction.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[2] R. Liu, J. Lehman, P. Molino, F. P. Such, E. Frank, A. Sergeev, and J. Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. *arXiv preprint arXiv:1807.03247*, 2018. 2

[3] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016. 2

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

Figure 4. Visualized examples in COFW test dataset. (Red denotes predicted values by ADNet model and Green denotes ground truth.)

Figure 5. Visualized examples in the 300W test dataset. (Red denotes predicted values by ADNet model and Green denotes ground truth.)

Figure 6. Visualized examples in the WFLW test dataset. (Red denotes predicted values by ADNet model and Green denotes ground truth.)

Components	Edges	Methods	Overall NME	Normal NME	Tangent NME	Bias Rate
-	Overall	Baseline	3.38	1.91	2.55	33.51%
		ADNet	2.93	1.54	2.28	48.05%
Contour	Face Contour	Baseline	5.85	2.97	4.73	59.20%
		ADNet	5.45	2.58	4.57	77.13%
Eyebrow	Right Eyebrow	Baseline	3.62	2.10	2.75	30.51%
		ADNet	3.31	1.86	2.56	37.35%
	Left Eyebrow	Baseline	3.44	1.99	2.62	31.62%
		ADNet	3.15	1.75	2.45	40.24%
Nose	Nose Middle Line	Baseline	2.13	1.78	1.59	35.13%
		ADNet	1.97	1.01	1.53	51.03%
	Nose Bottom Line	Baseline	2.31	1.43	1.66	15.59%
		ADNet	2.11	1.26	1.56	23.27%
Eye	Right Eye Superior Margin	Baseline	1.88	1.23	1.25	1.83%
		ADNet	1.48	0.94	1.01	7.85%
	Right Eye Inferior Margin	Baseline	1.81	1.19	1.22	2.52%
		ADNet	1.42	0.89	0.98	10.11%
	Left Eye Superior Margin	Baseline	1.83	1.20	1.22	1.65%
		ADNet	1.43	0.92	0.96	3.96%
	Left Eye Inferior Margin	Baseline	1.80	1.17	1.20	2.56%
		ADNet	1.39	0.87	0.94	8.00%
Mouth	Outer Lip Superior Margin	Baseline	2.35	1.48	1.64	10.80%
		ADNet	2.01	1.18	1.47	24.25%
	Outer Lip Inferior Margin	Baseline	2.81	1.69	2.06	21.89%
		ADNet	2.62	1.52	1.98	30.26%
	Inner Lip Superior Margin	Baseline	2.15	1.32	1.49	12.61%
		ADNet	1.79	0.97	1.33	37.37%
	Inner Lip Inferior Margin	Baseline	2.48	1.53	1.79	16.99%
		ADNet	2.13	1.24	1.64	32.25%

Table 7. Evaluation of individual edges on 300W.