Towards Accurate Facial Landmark Detection via Cascaded Transformers – Supplementary Material

Hui Li*[1], Zidong Guo*[1], Seon-Min Rhee[2], Seungju Han[2], Jae-Joon Han[2]
[1]Samsung R&D Institute China Xi'an (SRCX)
[2]Samsung Advanced Institute of Technology (SAIT), South Korea
hui01.li, zidong.guo, s.rhee, sj75.han, jae-joon.han@samsung.com

1. Introduction

This supplementary material provides the following information: Section 2 indicates detailed evaluation of our proposed landmark detector on WFLW subsets. Section 3 presents another group of cross-dataset evaluation results. Section 4 visualizes the deformable attention for each head. Section 5 presents more landmark detection results to show the advantage of our model. Section 6 presents some failure cases by our method for further improvement. All the models in this material use default configuration.

2. Experiments on WFLW

WFLW [15] consists of various challenges, *i.e.*, pose, expression, illumination, makeup, occlusion and blur. We conduct experiments on the full WFLW test set and six subsets, and compare with SOTA methods to show the superiority of our models. As presented in Table 1, DTLD+ and DTLD achieve the best and the second best separately, compared with other methods. Specifically, DTLD+ gets the lowest NME on WFLW-Full as well as the majority of subsets, except "illumination" where DTLD shows the lowest NME. DTLD+ improves greatly on the "expression" subset, surpassing the second best (PIPNet-101 [8]) with a relative increase of 8.3%, and the second best that adopts ResNet-18 as backbone (PIPNet-18) by 14.0%. In the subset of "occlusion", the advantage is relatively small (0.4% relative improve compared to the second best), indicating the potential research direction for further improvement.

3. More Cross-dataset Evaluation

Here we conduct another group of cross-dataset evaluation to prove the robustness of our model, following the experimental setting in [9]. To be specific, we train DTLD from scratch on the training data of 300W Split2, and evaluate it on 300W Split2 test data, Menpo frontal [5, 13, 17] and COFW68. There are 3837 images in 300W Split2 train set and 600 images in test set. The 6679 near-frontal training images in Menpo 2D (denoted as Menpo frontal) are adopted here for evaluation, as well as the 507 test images in COFW68. Following [9], here we adopt NME_{box} and AUC_{box} as the evaluation metrics. NME_{box} uses the geometric mean of the width and height of the ground-truth bounding box ($\sqrt{w_{bbox} \cdot h_{bbox}}$) as the normalization distance D. AUC (Area Under Curve) is computed as the area under the cumulative distribution curve, up to a cutoff NME value. The cumulative distribution curve is plotted by the fraction of test images whose NME is less than or equal to the specific NME value on the horizontal axis. Here we use NME_{box} and the cutoff value of 7%. The lower the NME_{box}, the higher the AUC_{box}, the better the performance. Experimental results are presented in Table 2. Our DTLD without any pretraining achieves the best detection accuracy on 300W Split2 test set and COFW68, even surpassing other methods pretrained on 300W-LP-2D [18]. On Menpo frontal, our DTLD is still better than previous models without pretraining.

4. Deformable Attention Visualization on Multi-Heads

We visualize the sampling points and deformable attention weights for each head in Figure 1. Specifically, for one head, we combine sampling points from different level of feature maps. The brighter the point, the greater the weight. Figure 1 is from the last DTLD decoder layer. The visualization illustrates intuitively that different heads will pay attention to different directions on image features and sampling points near the landmark to be detected gain more attention.

5. Landmark Detection Visualization

We present more samples to show the landmark prediction results of DTLD under different scenarios in WFLW subsets, such as different occlusion proportions and region-

*The first two authors equally contributed to this work. H.Li is the corresponding author.

Method	Year	Backbone	WFLW						
			Full	Pose	Expr.	Illu.	M.u.	Occ.	Blur
LAB [16]	2018	Hourglass	5.27	10.24	5.51	5.23	5.15	6.79	6.23
Wing [7]	2018	ResNet-50	4.99	8.43	5.21	4.88	5.26	6.21	5.81
DeCaFa [4]	2019	Cascaded U-Net	4.62	8.11	4.65	4.41	4.63	5.74	5.38
DAG [10]	2020	HRNet-W18	4.21	7.36	4.49	4.12	4.05	4.98	4.82
HRNet [12]	2019	HRNet-W18	4.60	7.94	4.85	4.55	4.29	5.44	5.42
AWing [14]	2019	Hourglass	4.36	7.38	4.58	4.32	4.27	5.19	4.96
AVS [11]	2019	ITN-CPM	4.39	8.42	4.68	4.24	4.37	5.60	4.86
PIPNet-18 [8]	2020	**ResNet-18**	4.57	8.02	4.73	4.39	4.38	5.66	5.25
PIPNet-101 [8]	2020	ResNet-101	4.31	7.51	4.44	4.19	4.02	5.36	5.02
DTLD	2021	**ResNet-18**	*4.08*	**7.06**	*4.20*	**3.96**	*3.92*	**4.96**	*4.75*
DTLD+	2021	**ResNet-18**	**4.05**	**7.06**	**4.07**	*4.02*	**3.83**	**4.96**	**4.74**

Table 1. Comparison with state-of-the-art methods on WFLW (Full set and six subsets). The results are in NME (%). Our DTLD+ achieves the best and DTLD achieves the second best in all sets with a simple ResNet-18 backbone.

Methods	NME_{box} (%) (↓)			AUC^{7}_{box} (%) (↑)		
	300W	Menpo	COFW68	300W	Menpo	COFW68
SAN* [3, 6]	2.86	2.95	3.50	59.7	61.9	51.9
2D-FAN* [1]	2.32	2.16	2.95	66.5	69.0	57.5
Softlabel* [3]	2.32	2.27	2.92	66.6	67.4	57.9
KDN [2]	2.49	2.26	–	67.3	68.2	–
KDN* [2]	2.21	**2.01**	2.73	68.3	71.1	60.1
LUVLi [9]	2.24	2.18	2.75	68.3	70.1	60.8
LUVLi* [9]	2.10	2.04	2.57	70.2	**71.9**	63.4
DTLD(s)	**2.05**	2.10	**2.47**	**70.9**	71.8	**65.0**

Table 2. Another group of cross-dataset evaluation. **DTLD(s)** is our proposed DTLD model trained from scratch. The methods marked with * are pretrained on 300W-LP-2D. Our DTLD exceeds previous models without pretraining and is even better than some models pretrained on 300W-LP-2D.

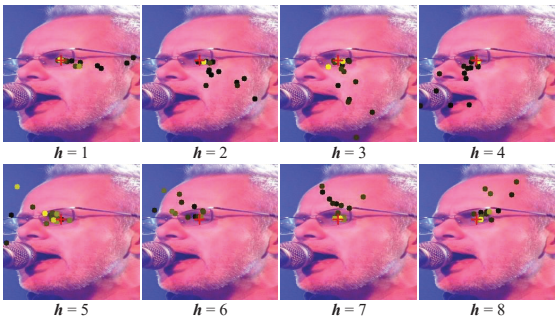

Figure 1. Visualizations of sampling points of the last DTLD decoder layer. The red cross denote the ground-truth, while others dots show the sampling points with attention weights expressed by colors. The brighter the point, the greater the weight. We combine the sampling points from all feature maps for each head.

s, various facial expressions, make-up, poses and illuminations, *et al*. Figure 2 indicates that our method can accurately predict the facial landmarks in different situations, even under extreme poses and makeup. The results also prove that the structural information among landmarks is well learned by our method.

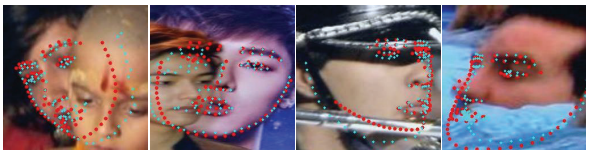

Figure 3. Visualizations of some typical failures. Red represents the ground truth, and cyan represents our predictions.

6. Failure Case Analysis

Although our model shows strong superiority on facial landmark detection, it is still weak for face image with severe occlusions, especially obscured by other people, as illustrated in Figure 3. Specifically, 1) If the challenge (*i.e*., blurring, occlusion, *etc*.) causes a great uncertainty on face boundary inference, our model may fail. 2) If the face to be aligned is obscured by another face, our model has difficulty in distinguishing the target character, thus leading to large errors. 3) The ambiguity of landmark annotations

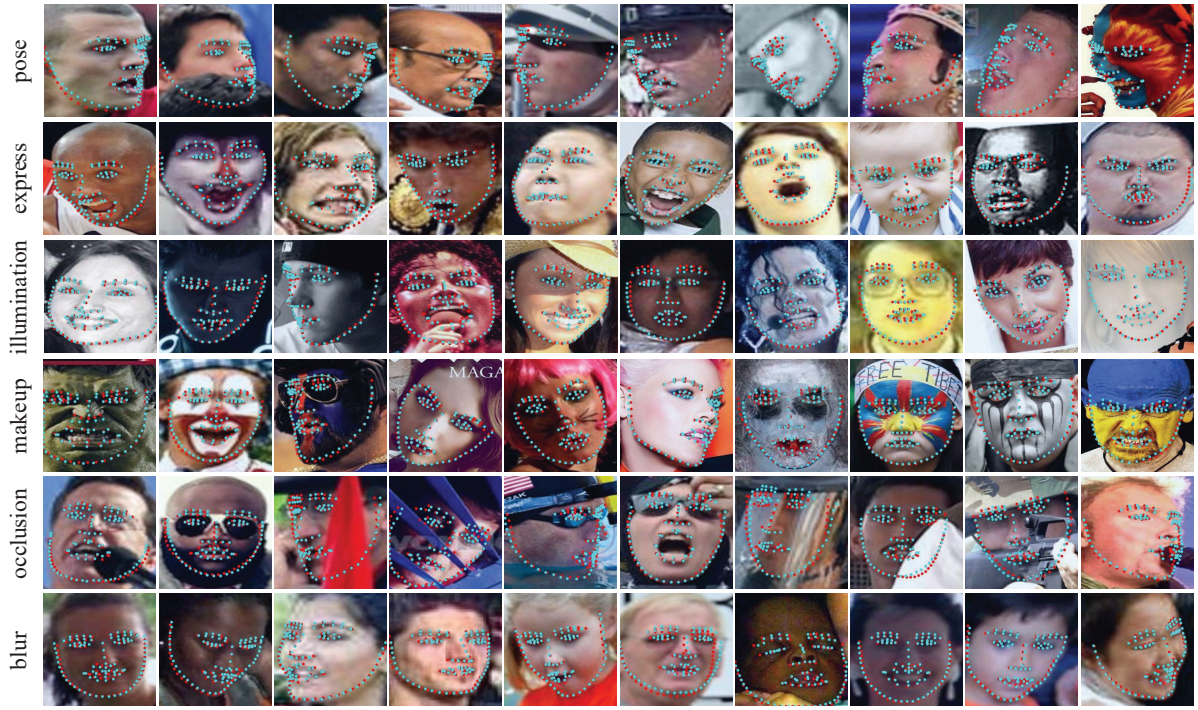

Figure 2. Typical samples from different WFLW subsets. Red denotes the ground truth, and cyan represents our predictions.

may lead to poor performance, especially for landmarks on face boundary. For these weaknesses, a possible solution is to make better use of the connections between landmarks to infer the invisible part. We leave it as a future work.

References

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

[2] Lisha Chen and Qiang Ji. Kernel density network for quantifying regression uncertainty in face alignment. In *NeurIPS*, 2018. 2

[3] Lisha Chen, Hui Su, and Qiang Ji. Face alignment with kernel density deep neural network. In *ICCV*, 2019. 2

[4] Arnaud Dapogny, Kevin Bailly, and Matthieu Cord. Decafa: Deep convolutional cascade for face alignment in the wild. In *ICCV*, pages 6893–6901, 2019. 2

[5] Jiankang Deng, Anastasios Roussos, Grigorios Chrysos, Evangelos Ververas, Irene Kotsia, Jie Shen, and Stefanos Zafeiriou. The menpo benchmark for multi-pose 2d and 3d facial landmark localisation and tracking. *IJCV*, (127):599–624, 2019. 1

[6] Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Style aggregated network for facial landmark detection. In *CVPR*, 2018. 2

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

[8] Haibo Jin, Shengcai Liao, and Ling Shao. Pixel-in-pixel net: Towards efficient facial landmark detection in the wild. *IJCV*, pages 1–21, 2021. 1, 2

[9] Abhinav Kumar, Tim K Marks, Wenxuan Mou, Ye Wang, Michael Jones, Anoop Cherian, Toshiaki Koike-Akino, Xiaoming Liu, and Chen Feng. Luvli face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood. In *CVPR*, pages 8236–8246, 2020. 1, 2

[10] Weijian Li, Yuhang Lu, Kang Zheng, Haofu Liao, Chihung Lin, Jiebo Luo, Chi-Tung Cheng, Jing Xiao, Le Lu, Chang-Fu Kuo, et al. Structured landmark detection via topology-adapting deep graph learning. In *ECCV*, pages 266–283. Springer, 2020. 2

[11] Shengju Qian, Keqiang Sun, Wayne Wu, Chen Qian, and Jiaya Jia. Aggregation via separation: Boosting facial landmark detector with semi-supervised style translation'. In *ICCV*, 2019. 2

[12] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, pages 5693–5703, 2019. 2

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

[14] 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

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

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

[17] Stefanos Zafeiriou, George Trigeorgis, Grigorios Chrysos, Jiankang Deng, and Jie Shen. The menpo facial landmark localisation challenge: A step closer to the solution. In *CVPRW*, pages 170–179, 2017. 1

[18] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Li. Face alignment across large poses A 3d solution. In *CVPR*, 2016. 1