Supplementary Material for Cross-domain Structured Landmark Detection via Progressive Topology-Adapting Deep Graph Learning

Weijian Li[1,2], Yuhang Lu[1,3], Kang Zheng[1], Haofu Liao[2], Chihung Lin[5], Jiebo Luo[2], Chi-Tung Cheng[5], Jing Xiao[4], Le Lu[1], Chang-Fu Kuo[5], and Shun Miao[1]

[1] PAII. Inc., Bethesda, MD, USA
[2] Department of Computer Science, University of Rochester, Rochester, NY, USA
[3] Department of Computer Science and Engineering, University of South Carolina, Columbia, SC, USA
[4] Ping An Technology, Shenzhen, China
[5] Chang Gung Memorial Hospital, Linkou, Taiwan, ROC

1 Additional discussions with related works:

Though some recent works [6,1,4] propose to model landmark relationship, our problem/method has large differences from them. Tompson *et al.* [6] propose to use spatial information in a post-processing step to filter outliers, while we leverage visual-spatial joint features for landmark regression. Also, the PAF proposed by Cao *et al.* [1] focuses on a different task of assembling detected key points for multi-person parsing. Zhao *et al.* [9] focus differently on predicting 3D poses from 2D joints. Their 2D joints are generated by a pre-trained 2D pose estimation network. Besides, their network structure is predefined by a fixed adjacency matrix while we actively learn the structures. Payer *et al.* [4], propose a spatial configuration branch to disambiguate candidates from the heatmap predictions. There is no explicit landmark structure modeling. In contrast, we explicitly model shape through a graph representation with learnable connectivity.

Among the SOTA, WING [2] is pure coordinate-based, while LAB [8] and AWING [7] integrate face boundary information via heatmap, which is their key contributions. The gap between WING and AWING is significant on WFLW, which is a more challenging dataset than 300W in terms of dataset scale, pose variations, occlusions, etc. Our method performs significantly better than WING on WFLW by reducing the failure rate by 50%, and is competitive to AWING. In addition, WING focuses on loss design, which is orthogonal and complementary to our novelty. By employing WING loss in our method, our performance can be further improved (e.g., on 300W, inter-pupil NME from 4.27 to 4.21 and inter-ocular NME from 3.04 to 3.01). While LAB and AWING utilize global representation, human knowledge on face structure via a boundary heatmap is injected, leading to task-specific solutions. In contrast, our method is a general landmark detection method to model the structural information via a self-learned graph structure.

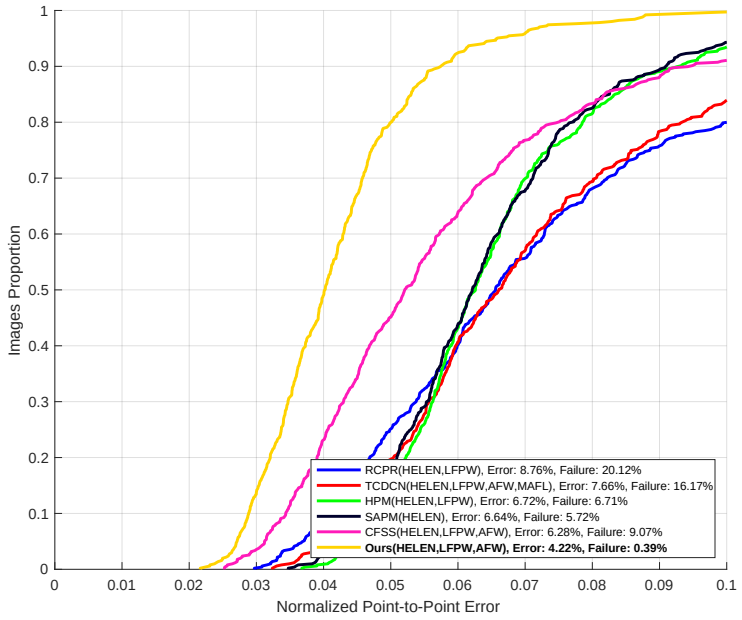

Fig. 1: Cumulative Errors Distribution (CED) curve results on the COFW-68 test set.

2 CED Curve:

Following previous works [8,5], we report Cumulative Errors Distribution (CED) curve result on cross-evaluations of COFW-68 test set. Recall that the success rate measures the proportion of images that have a localization error below a certain threshold [3]. Thus, given a range of thresholds, the corresponding success rates will form a distribution which is considered as Cumulative Error Distribution (CED). For clearer comparison, we include both Normalized Mean Error (**Error**) as well as the Failure Rate (i.e. $1 - SuccessRate$) (**Failure**) at threshold of 0.1. As we can see from Figure 1, our model outperforms previous methods by a large margin, especially in Failure Rate which is reduced to 0.39% for the first time. The comparison of numerical NME and Failure Rate values with the other state-of-the-arts can be found in Table 3 in our submitted ECCV-20 main paper.

3 Ablation Studies

Here we conduct three more types of ablation studies, namely: (1) The comparison of the transformation method used in GCN-global. (2) The effectiveness of the proposed GCN modules.(3) The comparison of different number of regression steps used in GCN-local. Results are recorded in Table 1.

Table 1: Ablation studies on the proposed model with 300W fullset under Inter-Ocular normalization.

Different Transformations	Affine Transformation	Perspective Transformation (Ours)
NME	3.13	**3.04**

Effectivenes of GCN modules	Replace GCN-global with CNN	Replace GCN-local with MLP
NME	3.12	3.18

Different GCN Steps	Step=1	Step=3 (Ours)	Step=5	Step=7
NME	3.24	**3.04**	3.07	3.11

Choice of transformations: We experiment two types of GCN-global choices: (1) Adopt Affine Transformation. In this case, the performance of our GCN-global module drops to 3.13.(2) Adopt Perspective Transformation. We achieve the best result as 3.04 which is also reported in our main paper. This indicates that GCN-global can better locates ROIs with the more flexible perspective transformation.

Effectiveness of GCN modules: We examine the effectiveness of the proposed GCN modules by: (1) Replacing GCN-global with a CNN block: we replace the GCN-global module with a 2-layer CNN (Conv/BN/ReLU) with Global Average Pooling predicting 9 transformation parameters. The average error increased from 3.04 to 3.12. (2) Replacing GCN-local with a MLP block: we remove the connectivity used in GCN-local, making it a simple MLP (FC/ReLU). The average error increased from 3.04 to 3.18. These indicating the importance of the proposed GCN modules.

Number of steps: We analyze different choices of steps for GCN-local. Results are shown in Table 1. The overall performance improves as the number of steps increases indicating the benefit of cascading multiple regressions. The best performance is achieved when GCN-local is implemented with three iterations.

4 More Settings:

We describe more settings for training the model. Adam optimizer is adopted with initial learning rate $lr = 0.0001$. The learning rate decreases at every 100 epochs. $L2$ penalty is applied to the training parameters with rate 0.0001. Margin for training GCN-global is set to $m = 0.1$ for Face300W, $m = 0.15$ for WFLW, $m = 0.15$ for three Medical datasets. All data augmentations we used: (1) Rotate input image with a random angle in [-30, 30]. (2) Random flip the input image horizontally. (3) Scale input image with a random factor in [0.75, 1.25].

References

1. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR. pp. 7291–7299 (2017)
2. Feng, Z.H., Kittler, J., Awais, M., Huber, P., Wu, X.J.: Wing loss for robust facial landmark localisation with convolutional neural networks. In: CVPR. pp. 2235–2245 (2018)
3. Ghiasi, G., Fowlkes, C.C.: Occlusion coherence: Detecting and localizing occluded faces. arXiv preprint arXiv:1506.08347 (2015)
4. Payer, C., Štern, D., Bischof, H., Urschler, M.: Integrating spatial configuration into heatmap regression based CNNs for landmark localization. MIA **54**, 207–219 (may 2019). https://doi.org/10.1016/j.media.2019.03.007
5. Qian, S., Sun, K., Wu, W., Qian, C., Jia, J.: Aggregation via separation: Boosting facial landmark detector with semi-supervised style translation. In: ICCV. pp. 10153–10163 (2019)
6. Tompson, J.J., Jain, A., LeCun, Y., Bregler, C.: Joint training of a convolutional network and a graphical model for human pose estimation. In: NeurIPS. pp. 1799–1807 (2014)
7. Wang, X., Bo, L., Fuxin, L.: Adaptive wing loss for robust face alignment via heatmap regression. In: ICCV. pp. 6971–6981 (2019)
8. Wu, W., Qian, C., Yang, S., Wang, Q., Cai, Y., Zhou, Q.: Look at boundary: A boundary-aware face alignment algorithm. In: CVPR. pp. 2129–2138 (2018)
9. Zhao, L., Peng, X., Tian, Y., Kapadia, M., Metaxas, D.N.: Semantic graph convolutional networks for 3d human pose regression. In: CVPR. pp. 3425–3435 (2019)