Adaptive Wing Loss for Robust Face Alignment via Heatmap Regression Supplementary Material

Xinyao Wang[1,2] Liefeng Bo[2] Li Fuxin[1]
[1]Oregon State University [2]JD Digits
{wangxiny, lif}@oregonstate.edu, {xinyao.wang3, liefeng.bo}@jd.com

This document is served as supplementary material for our ICCV 2019 submission *Adaptive Wing Loss for Robust Face Alignment via Heatmap Regression*. The document consists of five sections. Section 1 provides more implementation details on the CoordConv with boundary coordinates. Section 2 shows more detailed information about the datasets we used in experiments. We performed evaluation on the AFLW [13] dataset in Section 3. Additional ablation studies are shown in Section 4, and some facial landmark localization visualizations are included in Section 5.

1. Implementation Detail of CoordConv on Boundary Information

In addition to original CoordConv [11], we add two coordinate encoding channels with boundary information. A visualization of this process is shown in Figure 1

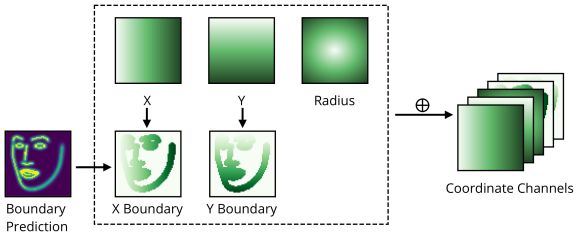

Figure 1: **CoodConv with Boundary Information**. X Boundary and Y Boundary are generated from X coordinate channel and Y coordinate channel respectively by a binary mask created from boundary prediction from the previous Hourglass module. The mask is generated by thresholding boundary prediction with a value of 0.05. (Best viewed in color).

2. Datasets Used in Our Experiments

The **COFW** [2] dataset includes 1,345 training images and 507 testing images annotated with 29 landmarks. This dataset is aimed to test the effectiveness of face alignment algorithms on faces with large pose and heavy occlusion. Various types of occlusions are introduced and result in a 23% occlusion on facial parts in average.

The **300W [20]** is widely used as a 2D face alignment benchmark with 68 annotated landmarks. 300W consists of the following subsets: LFPW [1], HELEN [10], AFW [26], XM2VTS [14] and an additional dataset with 135 images with large pose, occlusion and expressions called iBUG. To compare with other approaches, we adopt the widely used protocol described in [19] to train and evaluate our approach. More specifically, we use the training dataset of LFPW, HELEN, and the full AFW dataset as training dataset, and the test dataset of LFPW, HELEN and the full iBUG dataset as full test dataset. The full test dataset is then further split into two subsets, the test dataset of LFPW and HELEN is called the common test dataset, and iBUG is called the challenge test dataset. There is also a 300W private test dataset for the 300W contest, which contains 300 indoor and 300 outdoor faces. We also evaluated our approach on this dataset.

The **WFLW [22]** is a newly introduced dataset with 98 manually annotated landmarks that constitutes of 7,500 training images and 2,500 testing images. In addition to denser annotations, it also provides attribute annotations including pose, expression, illumination, make-up, occlusion and blur. The six different subsets can be used for analyzing algorithm performance on subsets with different properties separately. The WFLW is considered more difficult than commonly used datasets such as AFLW and 300W due to its more densely annotated landmarks and difficult faces with occlusion, blur, large pose, makeup, expression and illumination.

For the **LSP [8]** dataset, we used original label from author's official website[1][2]. Although images with original resolutions are also provided, we choose not to use them. Also, we did not use re-annotated labels on LSP extended 10,000 training images from [17]. Note that occluded keypoints are

[1]http://sam.johnson.io/research/lsp.html
[2]http://sam.johnson.io/research/lspet.html

annotated in LSP original dataset but not in LSP extended training dataset. During training, we did not calculate loss on occluded keypoints for LSP extended training dataset. During training and testing, we did not follow [16] to crop single person from images with multiple persons to retain the difficulties of this dataset. Data augmentations is performed similarly to training with face alignment datasets.

3. Evaluation on AFLW

The **AFLW [13]** dataset contains 24,368 faces with large poses. All faces are annotated by up to 21 landmarks per image, while the occluded landmarks were not labeled. For fair comparison with other methods we adopt the protocol from [24], which provides revised annotations with 19 landmarks. The training dataset contains 20,000 images, the full testing dataset contains 4,368 iamges. A subset of 1,314 frontal faces (no landmarks are occluded) are selected from the full test dataset as the frontal test set.

Method	Full(%)	Frontal(%)
$\text{RCPR}_{\text{CVPR 13}}$ [2]	3.73	2.87
$\text{ERT}_{\text{CVPR 14}}$ [9]	4.35	2.75
$\text{LBF}_{\text{CVPR 14}}$ [18]	4.25	2.74
$\text{CFSS}_{\text{CVPR 15}}$ [23]	3.92	2.68
$\text{CCL}_{\text{CVPR 16}}$ [25]	2.72	2.17
$\text{TSR}_{\text{CVPR 17}}$ [12]	2.17	-
$\text{DAC-OSR}_{\text{CVPR 17}}$ [6]	2.27	1.81
$\text{DCFE}_{\text{ECCV 18}}$ [21]	2.17	-
$\text{CPM+SBR}_{\text{CVPR 18}}$ [4]	2.14	-
$\text{SAN}_{\text{CVPR 18}}$ [3]	1.91	1.85
$\text{DSRN}_{\text{CVPR 18}}$ [15]	1.86	-
$\text{LAB}_{\text{CVPR 18}}$ [22]	1.85	1.62
$\text{Wing}_{\text{CVPR 18}}$ [5]	1.65	-
$\text{RCN}^{+}\text{(L+ELT+A)}_{\text{CVPR 18}}$ [7]	1.59	-
AWing(Ours)	**1.53**	**1.38**

Table 1: Mean error(%) on the AFLW testset

Evaluation results on the AFLW dataset are shown in Table 1. For AFLW dataset, we created boundary with a different scheme compared with Wu*et al*. [22] since insufficient landmarks are provided to generate all 14 boundary lines. We only use landmarks to generate left/right eyebrow, left/right eye line and noise bottom line. Even though we only have limited boundary information from 19 landmarks, our method is able to outperform the state-of-the-art methods in a large margin, which prove the robustness of our method to faces with large poses.

4. Additional Ablation Study

4.1. Effectiveness of Adaptive Wing loss on Training

Table 2 shows the effectiveness of our Adaptive Wing loss compare with MSE in terms of training loss w.r.t. the number of training epochs. Model trained with the Adaptive Wing loss is able to reduce the pixel-wise average MSE loss for almost 30%, and more than 23% on foreground pixels. Especially, this improvement comes at a mere 50 epochs, showing that the AWing loss improves convergence speed.

Loss \ Epoch	10	50	100	150	200
MSE_all	0.018	0.018	0.014	0.014	0.014
AW_all	0.018(-)	0.013(↓27%)	0.011(↓21%)	0.010(↓28%)	0.010(↓28%)
MSE_fg	1.17	1.25	0.95	0.94	0.92
AW_fg	1.13(↓3%)	0.87(↓30%)	0.74(↓22%)	0.72(↓23%)	0.71(↓23%)

Table 2: Training loss comparison. For fair comparison, the losses are evaluated with MSE. Model are trained with original stacked HG without weight map. Subscript _fg and _all stand for foreground pixels and all pixels respectively.

4.2. Robustness of Adaptive Wing loss on datasets with manually added annotation noise

We experimented our Adaptive Wing loss on the WFLW dataset with manually added labeling noise. The dataset is generated by randomly shifting $S\%$ of the inter-ocular distances from $P\%$ of the points with a random angle.

P(%)/S(%)	0/0	10/10	20/20	30/30
AWing	4.65	4.64	4.66	4.86

Table 3: AWing on the WFLW dataset with noise, without Weighted Loss Map, CoordConv and boundary.

4.3. Experiment on different number of HG stacks

We compare the performance of different number of stacks of HG module (see details in Table 4). With reduced number of HGs, the performance of our approach remains outstanding. Even with only one HG block, our approach still outperforms previous state-of-the-arts in all datasets except the common subset and the full dataset of 300W. Note that the one HG model is able to run at 120 FPS with Nvidia GTX 1080Ti graphics card. The result reflects the effectiveness of our approach on limited computation resources.

5. Result Visualization

For visualization purpose, some localization results are shown in Figure 2 and Figure 3

References

[1] Peter N Belhumeur, David W Jacobs, David J Kriegman, and Neeraj Kumar. Localizing parts of faces using a consensus of exemplars. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2930–2940, 2013.

[2] Xavier P Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust face landmark estimation under occlusion. In *Pro-*

	300W			300W	WFLW	COFW	GPU
	Common	Challange	Full	Private			Runtime (FPS)
Previous Best	3.27/2.90	7.18/5.15	4.04/3.35	3.88	5.11	5.27	-
AWing-1HG	3.89/2.81	6.80/4.72	4.46/3.18	3.74	4.50	5.18	120.47
AWing-2HGs	3.84/2.77	6.61/4.58	4.38/3.12	3.61	4.29	5.08	63.79
AWing-3HGs	3.79/2.73	6.61/4.58	4.34/3.10	3.59	4.24	5.01	45.29
AWing-4HGs	3.77/2.72	6.52/4.52	4.31/3.07	3.56	4.21	4.94	34.50

Table 4: **NME (%) on different number of stacks**. The NMEs of 300W are normalized by inter-pupil/inter-ocular distance, the NMEs of COFW are normalized by inter-pupil distance, and the NMEs of 300W Private and WFLW are normlaized by inter-ocular distance. NMEs in the "Previous Best" row are selected from Table 1 to 4 in our main paper. Runtime is evaluated on Nvidia GTX 1080Ti graphics card with batch size of 1.

ceedings of the IEEE International Conference on Computer Vision, pages 1513–1520, 2013.

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

[4] Xuanyi Dong, Shoou-I Yu, Xinshuo Weng, Shih-En Wei, Yi Yang, and Yaser Sheikh. Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 360–368, 2018.

[5] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[6] Zhen-Hua Feng, Josef Kittler, William Christmas, Patrik Huber, and Xiao-Jun Wu. Dynamic attention-controlled cascaded shape regression exploiting training data augmentation and fuzzy-set sample weighting. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3681–3690. IEEE, 2017.

[7] Sina Honari, Pavlo Molchanov, Stephen Tyree, Pascal Vincent, Christopher Pal, and Jan Kautz. Improving landmark localization with semi-supervised learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[8] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, volume 2, page 5, 2010.

[9] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, 2014.

[10] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S Huang. Interactive facial feature localization. In *European conference on computer vision*, pages 679–692. Springer, 2012.

[11] Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. *arXiv preprint arXiv:1807.03247*, 2018.

[12] Jiang-Jing Lv, Xiaohu Shao, Junliang Xing, Cheng Cheng, Xi Zhou, et al. A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In *CVPR*, volume 1, page 4, 2017.

[13] Peter M. Roth Martin Koestinger, Paul Wohlhart and Horst Bischof. Annotated Facial Landmarks in the Wild: A Large-scale, Real-world Database for Facial Landmark Localization. In *Proc. First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011.

[14] Kieron Messer, Jiri Matas, Josef Kittler, Juergen Luettin, and Gilbert Maitre. Xm2vtsdb: The extended m2vts database. In *Second international conference on audio and video-based biometric person authentication*, volume 964, pages 965–966, 1999.

[15] Xin Miao, Xiantong Zhen, Xianglong Liu, Cheng Deng, Vassilis Athitsos, and Heng Huang. Direct shape regression networks for end-to-end face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5040–5049, 2018.

[16] Xi Peng, Zhiqiang Tang, Fei Yang, Rogerio S Feris, and Dimitris Metaxas. Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2226–2234, 2018.

[17] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V Gehler, and Bernt Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4929–4937, 2016.

[18] Shaoqing Ren, Xudong Cao, Yichen Wei, and Jian Sun. Face alignment at 3000 fps via regressing local binary features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1685–1692, 2014.

[19] Shaoqing Ren, Xudong Cao, Yichen Wei, and Jian Sun. Face alignment via regressing local binary features. *IEEE Transactions on Image Processing*, 25(3):1233–1245, 2016.

[20] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 397–403. IEEE, 2013.

[21] Roberto Valle and M José. A deeply-initialized coarse-to-fine ensemble of regression trees for face alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 585–601, 2018.

[22] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[23] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang. Face alignment by coarse-to-fine shape searching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4998–5006, 2015.

[24] Shizhan Zhu, Cheng Li, Chen-Change Loy, and Xiaoou Tang. Unconstrained face alignment via cascaded compositional learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3409–3417, 2016.

[25] Shizhan Zhu, Cheng Li, Chen-Change Loy, and Xiaoou Tang. Unconstrained face alignment via cascaded compositional learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3409–3417, 2016.

[26] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2879–2886. IEEE, 2012.

Figure 2: **Result visualization 1**. Row 1-2: AFLW dataset, row 3-4: COFW dataset, row 5-6: 300W dataset.

Figure 3: **Result visualization 2**. Row 1-2: 300W private dataset, row 3-4: WFLW dataset.