FG-Net: Facial Action Unit Detection with Generalizable Pyramidal Features Supplementary Material

Yufeng Yin[1]*, Di Chang[1], Guoxian Song[2], Shen Sang[2], Tiancheng Zhi[2],
Jing Liu[2], Linjie Luo[2], Mohammad Soleymani[1]
[1] University of Southern California, [2] ByteDance
{yufengy, dichang, msoleyma}@usc.edu,
{guoxiansong, shen.sang, tiancheng.zhi, jing.liu, linjie.luo}@bytedance.com

Region of Interest (ROI) for each AU. Following the previous work [21], we select two points on the face based on the most representative landmarks. Detailed postions are shown in Table 1. Note that the ROI locations of AU4, AU9, and AU26 are different from Zhang *et al*. [21] for better identification.

Table 1. Region of Interest (ROI) for each action unit (AU). Scale is measured by inner-ocular distance (IOD). Landmark (Lmk) positions are illustrated in Figure 1.

AU	Description	ROI Center
1	Inner Brow Raiser	Lmk 22, 23
2	Outer Brow Raiser	Lmk 18, 27
4	Brow Lowerer	Brow center
6	Cheek Raiser	1 scale below eye center
7	Lid Tightener	Lmk 39, 44
9	Nose Wrinkler	Lmk 40, 43
10	Upper Lip Raiser	Lmk 51, 53
12	Lip Corner Puller	Lmk 49, 55
14	Dimpler	Lmk 49, 55
15	Lip Corner Depressor	Lmk 49, 55
17	Chin Raiser	0.5 scale below Lmk 57, 59
23	Lip Tightener	Lmk 52, 58
24	Lip Pressor	Lmk 52, 58
25	Lips part	Lmk 52, 58
26	Jaw Drop	Lmk 57, 59

Detailed Results for Within-Domain AU Detection. We provide detailed within-domain evaluations for every individual AU on DISFA, BP4D in Table 2. The results show that FG-Net achieves competitive performance compared to the state-of-the-art which demonstrate that the pixel-wise features extracted from StyleGAN2 are beneficial for heatmap-based AU detection.

AU Intensity Estimation. Unlike AU detection which is a binary classification problem, intensity estimation provides a discrete regression from input face images. FG-Net addresses the AU detection problem using a heatmap regression which can be extended to AU intensity estimation.

*Work done during internship at ByteDance.

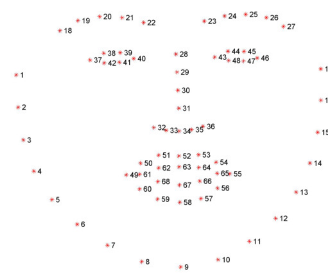

Figure 1. The positions for the 68 facial landmarks. Image is adapted from the iBUG 300-W dataset [9].

Specifically, following [8], the peak value for the heatmap is the corresponding AU intensity (ranging from 0 to 5) and we take the maximum of each heatmap as the predicted AU intensity.

The results on DISFA are shown in Table 3. The evaluation metrics are mean squared error (MSE ↓) and mean absolute error (MAE ↓). The results show that FG-Net also achieves competitive performance compared to the state-of-the-art for AU intensity estimation.

References

[1] Yanan Chang and Shangfei Wang. Knowledge-driven self-supervised representation learning for facial action unit recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20417–20426, 2022. 2

[2] Yingruo Fan, Jacqueline Lam, and Victor Li. Facial action unit intensity estimation via semantic correspondence learning with dynamic graph convolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12701–12708, 2020. 2

[3] Geethu Miriam Jacob and Bjorn Stenger. Facial action unit detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7680–7689, 2021. 2

Table 2. Within-domain evaluation in terms of F1 score (↑). Except for GH-Feat and ME-GraphAU + FFHQ pre-train, all the baseline numbers are from the original papers. Our method has competitive performance compared to the state-of-the-art.

(a) Within-domain evaluation on DISFA [7].

Methods	AU1	AU2	AU4	AU6	AU9	AU12	AU25	AU26	**Avg.**
DRML [22]	17.3	17.7	37.4	29.0	10.7	37.7	38.5	20.1	26.7
IdenNet [15]	25.5	34.8	64.5	45.2	44.6	70.7	81.0	55.0	52.6
SRERL [4]	45.7	47.8	59.6	47.1	45.6	73.5	84.3	43.6	55.9
UGN-B [12]	43.3	48.1	63.4	49.5	48.2	72.9	90.8	59.0	60.0
HMP-PS [13]	38.0	45.9	65.2	50.9	50.8	76.0	93.3	**67.6**	61.0
FAT [3]	46.1	48.6	72.8	**56.7**	50.0	72.1	90.8	55.4	61.5
Zhang *et al*. [21]	55.0	63.0	**74.6**	45.3	35.2	75.3	93.5	54.4	62.0
JÂA-Net [11]	62.4	60.7	67.1	41.1	45.1	73.5	90.9	67.4	63.5
PIAP [14]	50.2	51.8	71.9	50.6	54.5	**79.7**	94.1	57.2	63.8
Chang *et al*. [1]	60.4	59.2	67.5	52.7	51.5	76.1	91.3	57.7	64.5
ME-GraphAU [6]	54.6	47.1	72.9	54.0	**55.7**	76.7	91.1	53.0	63.1
ME-GraphAU + FFHQ pre-train	46.1	44.8	72.4	48.2	48.1	70.3	90.9	55.4	59.5
GH-Feat [17]	16.9	13.8	39.1	37.1	16.7	65.0	78.7	28.1	36.9
Ours	**63.6**	**66.9**	72.5	50.7	48.8	76.5	**94.1**	50.1	**65.4**

(b) Within-domain evaluation on BP4D [18].

Methods	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU14	AU15	AU17	AU23	AU24	**Avg.**
DRML [22]	36.4	41.8	43.0	55.0	67.0	66.3	65.8	54.1	33.2	48.0	31.7	30.0	48.3
IdenNet [15]	50.5	35.9	50.6	77.2	74.2	82.9	85.1	63.0	42.2	60.8	42.1	46.5	59.3
SRERL [4]	46.9	45.3	55.6	77.1	78.4	83.5	87.6	63.9	52.2	63.9	47.1	53.3	62.9
UGN-B [12]	54.2	46.4	56.8	76.2	76.7	82.4	86.1	64.7	51.2	63.1	48.5	53.6	63.3
HMP-PS [13]	53.1	46.1	56.0	76.5	76.9	82.1	86.4	64.8	51.5	63.0	49.9	54.5	63.4
FAT [3]	51.7	**49.3**	61.0	77.8	79.5	82.9	86.3	67.6	51.9	63.0	43.7	**56.3**	64.2
Zhang *et al*. [21]	52.6	47.0	**61.4**	76.8	79.2	83.5	88.6	60.4	49.3	62.6	**50.8**	49.6	63.5
JÂA-Net [11]	53.8	47.8	58.2	78.5	75.8	82.7	88.2	63.7	43.3	61.8	45.6	49.9	62.4
PIAP [14]	**54.2**	47.1	54.0	79.0	78.2	**86.3**	**89.5**	66.1	49.7	63.2	49.9	52.0	64.1
Chang *et al*. [1]	53.3	47.4	56.2	79.4	**80.7**	85.1	89.0	67.4	**55.9**	61.9	48.5	49.0	64.5
ME-GraphAU [6]	52.7	44.3	60.9	**79.9**	80.1	85.3	89.2	**69.4**	55.4	64.4	49.8	55.1	**65.5**
ME-GraphAU + FFHQ pre-train	51.1	38.8	57.0	76.8	78.9	83.2	88.3	64.1	44.0	61.5	44.5	45.2	61.1
GH-Feat [17]	42.7	43.2	47.6	73.5	66.2	75.6	83.8	54.2	43.9	62.4	41.9	45.0	56.7
Ours	52.6	48.8	57.1	79.8	77.5	85.6	89.3	68.0	45.6	**64.8**	46.2	55.7	64.3

Table 3. AU intensity estimation on DISFA [7] in terms of MSE (↓) and MAE (↓). Our method has competitive performacne compared to the state-of-the-art.

Metric	Method	AU1	AU2	AU4	AU5	AU6	AU9	AU12	AU15	AU17	AU20	AU25	AU26	**Avg.**
MSE	2DC [5]	0.32	0.39	0.53	0.26	0.43	0.30	**0.25**	0.27	0.61	0.18	0.37	0.55	0.37
	HR [8]	0.41	0.37	0.70	0.08	0.44	0.30	0.29	0.14	0.26	**0.16**	0.24	0.39	0.32
	APs [10]	0.68	0.59	**0.40**	**0.03**	0.49	**0.15**	0.26	**0.13**	**0.22**	0.20	0.35	**0.17**	0.30
	Ours	**0.25**	**0.21**	0.47	0.07	**0.35**	0.20	0.27	0.15	0.27	**0.16**	**0.23**	0.40	**0.25**
MAE	KJRE [20]	1.02	0.92	1.86	0.70	0.79	0.87	0.77	0.60	0.80	0.72	0.96	0.94	0.91
	CCNN-IT [16]	0.73	0.72	1.03	0.21	0.72	0.51	0.72	0.43	0.50	0.44	1.16	0.79	0.66
	KBSS [19]	0.48	0.49	0.57	0.08	0.26	0.22	0.33	0.15	0.44	0.22	0.43	0.36	0.33
	SCC-Heatmap [2]	**0.16**	**0.16**	**0.27**	**0.03**	**0.25**	**0.13**	**0.32**	**0.15**	**0.20**	**0.09**	**0.30**	**0.32**	**0.20**
	Ours	0.19	**0.16**	0.36	0.06	0.31	0.17	**0.32**	0.18	0.27	0.15	0.34	0.41	0.25

[4] Guanbin Li, Xin Zhu, Yirui Zeng, Qing Wang, and Liang Lin. Semantic relationships guided representation learning for facial action unit recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8594–8601, 2019. 2

[5] Dieu Linh Tran, Robert Walecki, Stefanos Eleftheriadis, Bjorn Schuller, Maja Pantic, et al. Deepcoder: Semi-parametric variational autoencoders for automatic facial action coding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3190–3199, 2017. 2

[6] Cheng Luo, Siyang Song, Weicheng Xie, Linlin Shen, and Hatice Gunes. Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition. *arXiv preprint arXiv:2205.01782*, 2022. 2

[7] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013. 2

[8] Ioanna Ntinou, Enrique Sanchez, Adrian Bulat, Michel Valstar, and Yorgos Tzimiropoulos. A transfer learning approach to heatmap regression for action unit intensity estimation. *IEEE Transactions on Affective Computing*, 2021. 1, 2

[9] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 397–403, 2013. 1

[10] Enrique Sanchez, Mani Kumar Tellamekala, Michel Valstar, and Georgios Tzimiropoulos. Affective processes: stochastic modelling of temporal context for emotion and facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9074–9084, 2021. 2

[11] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. Jaanet: Joint facial action unit detection and face alignment via adaptive attention. *International Journal of Computer Vision*, 129(2):321–340, 2021. 2

[12] Tengfei Song, Lisha Chen, Wenming Zheng, and Qiang Ji. Uncertain graph neural networks for facial action unit detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5993–6001, 2021. 2

[13] Tengfei Song, Zijun Cui, Wenming Zheng, and Qiang Ji. Hybrid message passing with performance-driven structures for facial action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6267–6276, 2021. 2

[14] Yang Tang, Wangding Zeng, Dafei Zhao, and Honggang Zhang. Piap-df: Pixel-interested and anti person-specific facial action unit detection net with discrete feedback learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12899–12908, 2021. 2

[15] Cheng-Hao Tu, Chih-Yuan Yang, and Jane Yung-jen Hsu. Idennet: Identity-aware facial action unit detection. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8. IEEE, 2019. 2

[16] Robert Walecki, Vladimir Pavlovic, Björn Schuller, Maja Pantic, et al. Deep structured learning for facial action unit intensity estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3405–3414, 2017. 2

[17] Yinghao Xu, Yujun Shen, Jiapeng Zhu, Ceyuan Yang, and Bolei Zhou. Generative hierarchical features from synthesizing images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4432–4442, 2021. 2

[18] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014. 2

[19] Yong Zhang, Weiming Dong, Bao-Gang Hu, and Qiang Ji. Weakly-supervised deep convolutional neural network learning for facial action unit intensity estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2314–2323, 2018. 2

[20] Yong Zhang, Baoyuan Wu, Weiming Dong, Zhifeng Li, Wei Liu, Bao-Gang Hu, and Qiang Ji. Joint representation and estimator learning for facial action unit intensity estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3457–3466, 2019. 2

[21] Zheng Zhang, Taoyue Wang, and Lijun Yin. Region of interest based graph convolution: A heatmap regression approach for action unit detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2890–2898, 2020. 1, 2

[22] Kaili Zhao, Wen-Sheng Chu, and Honggang Zhang. Deep region and multi-label learning for facial action unit detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3391–3399, 2016. 2