Supplementary Material for "Context-aware Feature and Label Fusion for Facial Action Unit Intensity Estimation with Partially Labeled Data"

Yong Zhang[1]*, Haiyong Jiang[2]*, Baoyuan Wu[1]†, Yanbo Fan[1], Qiang Ji[3]
[1]Tencent AI Lab, [2]Nanyang Technological University, Singapore, [3]Rensselaer Polytechnic Institute

1. Locations of Local patches

In Sec. 4.1 of the main script, we briefly presente the size of local patches. Here, we present more details about the preprocessing on local patches. Faces in sequences are cropped out and aligned by using two eye centers according to the provided facial landmarks in FERA 2015 and DISFA. Faces are resized to the size of 256×256. The coordinates of two eye centers are (78, 80) and (178, 80). In this work, we extract 8 local patches around facial components since AUs are closely associated with facial components (see Fig. 1). Each patch involves multiple AUs. The coordinates of patches are as follows: (row,col,width,height)

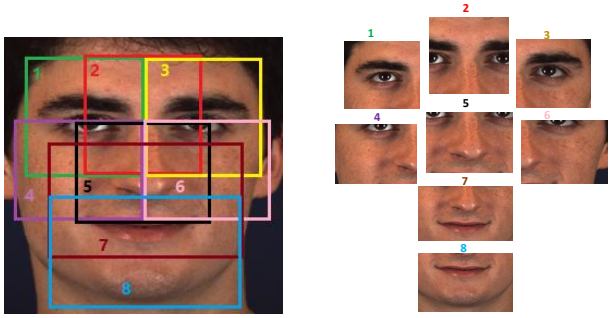
- Patch 1: (20,20,108,108)
- Patch 2: (20,74,108,108)
- Patch 3: (20,128,108,108)
- Patch 4: (95,8,120,80)
- Patch 5: (95,53,150,80)
- Patch 6: (95,128,120,80)
- Patch 7: (110,40,176,100)
- Patch 8: (150,40,176,100)

Each patch is resized to the sized of 32×32. We feed each patch into an individual ResNet18 to extract local features as each patch contains its own facial patterns when performing a meaningful expression. The whole face is resized to the size of 32×32 and is fed to a ResNet18 to extract global features.

2. Details about the Network

In Sec. 3.1 of the main script, for the feature fusion module, we use ResNet18 [2] to extract features for local patches and the whole face. The output dimension of ResNet18 is 256, *i.e.,* the dimension of $\mathbf{h}_m$ is 256. The fused feature vector is the concatenation of global features and local features. Its dimension is 512, *i.e.,* $d_f = 512$. The task-related context is denoted as $\mathbf{C} \in \mathcal{R}^{K \times d_c}$, where $d_c = 256$. K is 5 in FERA 2015 and is 12 in DISFA.

In Sec. 3.2 of the main script, we use a one-layer LSTM network [3] to model temporal dynamics of AUs by predicting label attention. The dimension of LSTM input is $d_f + d_c$, *i.e.,* 768D. The dimension of LSTM output is 2. The dimension of the hidden layer of LSTM is 256.



(a) (b)

Figure 1. (a) Locations of patches. (b) Patches

3. Detailed Comparison with State-of-the-art Supervised Learning Methods

In Table 4 of the main script, we present the average performance of state-of-the-art supervised learning methods and our method. Here, we present the performance of each AU in Table 1. Our method uses only the intensity annotations of key frames in sequences while other methods use all the frames in sequences.

References

[1] A. Gudi, H. E. Tasli, T. M. Den Uyl, and A. Maroulis. Deep learning based facs action unit occurrence and intensity estimation. In *FG*, 2015. 2

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1, 2

*Authors contributed equally
†Corresponding author

Table 1. Comparison with state-of-the-art supervised learning methods. (*) Indicates results taken from the reference.

Database		FERA 2015						DISFA												
	AU	6	10	12	14	17	Avg	1	2	4	5	6	9	12	15	17	20	25	26	Avg
ICC	HBN [7]*	.760	.710	.850	.520	**.690**	**.706**	-	-	-	-	-	-	-	-	-	-	-	-	-
	Heatmap [4]*	**.790**	**.800**	**.860**	**.540**	.430	.684	-	-	-	-	-	-	-	-	-	-	-	-	-
	2DC [5]*	.760	.710	.850	.450	.530	.660	**.700**	**.550**	**.690**	.050	**.590**	**.570**	**.880**	**.320**	.100	.080	**.900**	.500	**.494**
	CCNN-IT [6]*	.750	.690	**.860**	.400	.450	.630	.200	.120	.460	.080	.480	.440	.730	.290	**.450**	**.210**	.600	.460	.377
	CNN [1]	.740	.653	.833	.223	.532	.596	.057	.040	.378	.158	.485	.331	.769	.206	.197	.117	.758	.436	.328
	ResNet18 [2]	.714	.634	.812	.283	.456	.580	.005	-.002	.323	.093	.426	.199	.713	.077	.146	.078	.712	.472	.270
	CFLF (ours)	.766	.703	.827	.411	.600	.661	.263	.194	.459	**.354**	.516	.356	.707	.183	.340	.206	.811	**.510**	.408
MAE	HBN [7]*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	Heatmap [4]*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	2DC [5]*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	CCNN-IT [6]*	1.170	1.430	.970	1.650	1.080	1.260	.730	.720	1.030	.210	.720	.510	.720	.430	.500	.440	1.160	.790	.663
	CNN [1]	.709	.874	.640	1.145	.717	.817	.532	.487	.670	.176	.354	.290	**.390**	.241	.466	.270	.670	.531	.423
	ResNet18 [2]	.743	.888	.752	1.247	.777	.881	.572	.433	.912	.183	.413	.446	.481	.303	.465	.302	.723	.556	.482
	CFLF (ours)	**.624**	**.830**	**.624**	**1.000**	**.626**	**.741**	**.326**	**.280**	**.605**	**.126**	**.350**	**.275**	.425	**.180**	**.290**	**.164**	**.530**	**.398**	**.329**

[3] S. Hochreiter and S. Jrgen. Long short-term memory. In *Neural computation*, 1997. 1

[4] E. Sanchez, G. Tzimiropoulos, and M. Valstar. Joint action unit localisation and intensity estimation through heatmap regression. *BMVC*, 2018. 2

[5] D. L. Tran, R. Walecki, O. Rudovic, S. Eleftheriadis, B. Schuller, and M. Pantic. Deepcoder: Semi-parametric variational autoencoders for automatic facial action coding. *ICCV*, 2017. 2

[6] R. Walecki, O. Rudovic, V. Pavlovic, B. Schuller, and M. Pantic. Deep structured learning for facial action unit intensity estimation. In *CVPR*, 2017. 2

[7] S. Wang, L. Hao, and Q. Ji. Facial action unit recognition and intensity estimation enhanced through label dependencies. *TIP*, 2019. 2