Supplementary Material: Self-Supervised Learning for Facial Action Unit Recognition through Temporal Consistency

Liupei Lu
liupeilu@usc.edu
Leili Tavabi
ltavabi@usc.edu
Mohammad Soleymani
soleymani@ict.usc.edu

Institute for Creative Technologies
University of Southern California
Playa Vista, CA
USA

We include detailed datasets pre-processing in Section 1, complete ablation study results in Section 2, AU descriptors in Section 3 and more image Retrieval Examples in Section 4.

1 Datasets

BP4D [9] has 41 subjects with around 146,000 AU labeled images. DISFA [6] contains 26 subjects with around 130,000 AU labeled frames. In DISFA, the AU intensity ranges from 0 to 5, and we followed the same procedure as [5, 8] by labeling frames with intensity greater than 1 to positive and the rest to negative. Similar to [8], both BP4D and DISFA images are cropped and resized into 256×256 and aligned with the dataset's provided facial landmarks. The AU detection is evaluated using a subject-independent 3-fold cross-validation, where two folds are used for training and the remaining one is used for testing.

EmotioNet [2] consists of 943,673 annotated images, each face being detected and cropped using dlib [3], resized and center-cropped into 256×256, similar to VoxCeleb2. We randomly split the dataset in 80/20 ratio with 754,938 images for training and 188,735 images for validation, and an independent set of about 25,000 images for testing. AffectNet contains eight classes of facial expressions (neutral, happy, sad, surprise, fear, disgust, anger, contempt). Similar to [4], a subset of AffectNet [7] is used as train/validation set with 85/15 ratio split, which provides around 287,000 images for training and 58,000 for testing. A separate set is used for validation with 4,000 images.

2 Additional ablation study results

Additional ablation results with per-AU and per-expression performances are available in Tables 1, 2, 3 and 4. Ablation studies on different network architures can be found in Table 5.

Method	1	2	4	6	7	10	12	14	15	17	23	24	avg
Triplet	27.6	24.0	38.3	73.6	70.2	79.3	80.7	56.7	25.9	49.5	23.7	28.1	48.1
N-pair-MC	21.8	15.0	28.8	60.8	65.0	73.6	69.7	50.7	32.9	42.3	21.7	20.5	41.9
Ranking k=1	35.2	**25.5**	30.2	71.3	69.6	**81.3**	83.3	59.1	30.3	56.1	27.0	33.4	50.2
Ranking k=2	34.2	25.5	40.3	73.2	66.1	80.8	**84.6**	58.5	25.7	53.6	28.1	27.6	49.9
Ranking k=4	39.5	25.3	40.8	**74.9**	**70.4**	77.0	82.6	**62.3**	**33.7**	44.4	24.5	32.6	50.7
Ranking Multi	**42.3**	24.3	**44.1**	71.8	67.8	77.6	83.3	61.2	31.6	**51.6**	**29.8**	**38.6**	**52.0**

Table 1: Detailed ablation result on BP4D dataset (F1-score).

Method	1	2	4	6	9	12	25	26	avg
Triplet	13.2	16.7	32.0	32.7	22.5	47.2	57.8	16.9	29.9
N-pair-MC	10.7	2.5	29.7	19.0	13.2	23.9	42.0	6.9	18.5
Ranking k = 1	10.8	20.7	43.3	37.6	12.2	68.7	62.9	**46.2**	37.8
Ranking k = 2	**36.2**	**41.7**	27.8	**38.1**	**23.6**	**69.3**	66.0	21.0	**40.5**
Ranking k = 4	21.6	27.1	**38.3**	24.7	21.8	57.2	58.7	22.7	34.0
Ranking Multi	18.7	27.4	35.1	33.6	20.7	67.5	**68.0**	43.8	39.4

Table 2: Detaield ablation results for DISFA dataset (F1-score).

Method	1	2	4	5	6	9	12	17	20	25	26	avg
Triplet	65.9	65.4	76.2	72.4	85.1	76.2	87.2	67.4	77.0	79.4	66.6	74.4
N-pair-MC	62.8	53.9	64.6	58.2	74.0	58.6	74.4	56.5	62.9	65.9	59.2	62.8
Ranking k = 1	68.1	71.4	78.5	76.2	91.5	80.0	94.7	71.8	75.4	84.0	69.3	78.3
Ranking k = 2	68.5	70.9	77.4	78.5	89.8	77.8	93.3	68.9	80.8	81.8	70.0	78.0
Ranking k = 4	68.6	70.9	78.4	78.6	89.0	79.7	92.2	66.5	77.3	81.0	68.3	77.3
Ranking Multi	**70.7**	**73.3**	**80.5**	**82.1**	**92.1**	**84.3**	**95.9**	**73.4**	**81.6**	**87.4**	**72.2**	**81.2**

Table 3: Detailed ablation results on EmotioNet dataset (AUC).

Method	Neutral	Happy	Sad	Surprise	Fear	Disgust	Anger	Contempt	avg.
Triplet	66.8	77.2	67.0	75.5	75.6	63.2	68.3	67.6	70.2
N-pair-MC	58.4	64.9	59.6	56.9	62.4	54.7	58.9	58.4	59.3
Ranking k = 1	70.6	90.0	70.8	76.7	78.2	73.6	75.2	74.2	76.2
Ranking k = 2	71.6	87.9	71.3	77.5	77.7	71.1	74.6	72.5	75.5
Ranking k = 4	70.0	85.5	68.7	76.9	78.9	70.7	73.1	70.8	74.3
Ranking Multi	**73.4**	**91.4**	**75.5**	**79.9**	**81.6**	**74.2**	**76.5**	**76.0**	**78.6**

Table 4: Detailed ablation results on AffectNet dataset (AUC).

Method	1	2	4	6	7	10	12	14	15	17	23	24	avg
MobileNet_v2	32.2	25.6	34.9	**72.8**	70.4	81.0	**85.4**	54.1	**33.0**	49.1	**32.0**	31.1	50.1
ResNet18	35.2	25.5	30.2	71.3	69.6	**81.3**	83.3	59.1	30.3	**56.1**	27.0	**33.4**	50.2
DenseNet121	**38.8**	**30.1**	**36.2**	70.8	**70.8**	81.1	85.3	**60.8**	29.9	51.2	29.5	32.7	**51.4**

Table 5: Detailed ablation result on BP4D dataset with different network architectures(F1-score) at k=1.

3 Action Unit descriptors

We listed explicit definition for each Action Unit (AU), that are mentioned in this paper, in Table 6

AU	Description	AU	Description	AU	Description
1	Inner brow raiser	9	Nose wrinkler	20	Lip stretcher
2	Outer brow raiser	10	Upper lip raiser	23	Lip tightener
4	Brow lowerer	12	Lip corner puller	24	Lip pressor
5	Upper lid raiser	14	Dimpler	25	Lips part
6	Cheek raiser	15	Lip corner depressor	26	Jaw drop
7	Lid tightener	17	Chin raiser		

Table 6: The definition of the Facial Action Units (AU), mentioned in this paper, according to FACS [1]

4 Image Retrieval Examples

Additional examples of retrieved images for our representations and Fab-Net [4] are provided in Fig. 1.

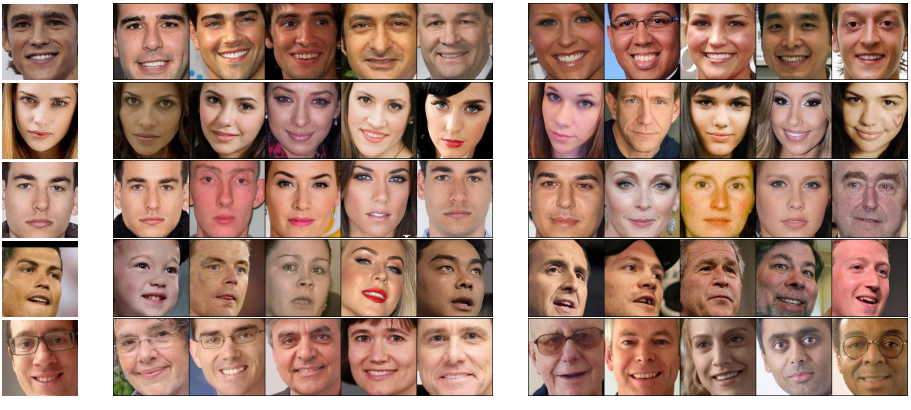

Figure 1: Additional Image Retrieval Examples: The top 5 image retrieval results for the query image (the single frame on the left). The left column shows our results, and Fab-Net retrieval results are on the right.

References

[1] Paul Ekman. Facial action coding system, 1977.

[2] C. Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M. Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[3] Davis King. Dlib c++ library. *Access on: http://dlib. net*, 2012.

[4] A Sophia Koepke, Olivia Wiles, and Andrew Zisserman. Self-supervised learning of a facial attribute embedding from video. In *BMVC*, page 302, 2018.

[5] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Self-supervised representation learning from videos for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10924–10933, 2019.

[6] 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.

[7] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.

[8] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. Deep adaptive attention for joint facial action unit detection and face alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 705–720, 2018.

[9] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, and Peng Liu. A high-resolution spontaneous 3d dynamic facial expression database. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6. IEEE, 2013.