Supplementary material for "FAN-Trans: Online Knowledge Distillation for Facial Action Unit Detection"

Jing Yang [1]
y.jing2016@gmail.com

Jie Shen [2]
jie.shen07@imperial.ac.uk

Yiming Lin [2]
yl1915@ic.ac.uk

Yordan Hristov
yshristov@gmail.com

Maja Pantic[2]
maja.pantic@gmail.com

[1]University of Nottingham, UK
[2]Imperial College London, UK

1. Convolution module

Following FAN [1], we utilize the multi-scale residual block in Convolution module. This block is superior to the basic block in capturing multi-scale feature representations. For the $768 \times 64 \times 64$ input feature, we first use 1×1 convolution to reduce channel dimension to 256. Then four multi-scale blocks equipped with max pooling layers are followed. The multi-scale block is shown in Figure 7.

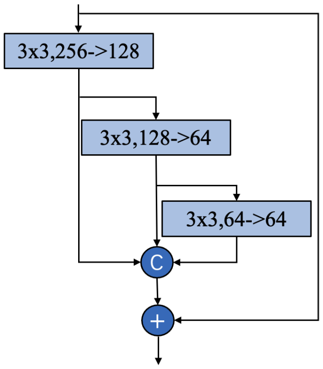

Figure 7. Multi-scale residual block in FAN [1].

2. Cross dataset testings

Unfortunately no recent work except of JÂANet (IJCV2021) is open source (or released models trained on BP4D or DISFA). Hence we could only compare the cross-dataset performance of our model with that of JÂANet, in the setting of training on BP4D and testing on DISFA. As shown in In Table 7, our method greatly outperforms JÂANet on all AUs in this setting.

Methods	AU1	AU2	AU4	AU6	AU12	AVG
JÂANet	19.2	16.1	28.5	30.6	35.4	26.0
FAN-Trans	**25.2**	**19.7**	**40.1**	**35.0**	**45.5**	33.1

Table 7. Evaluate JÂANet and FAN-Tran trained from BP4D on DISFA.(F1-score in %)

3. Complexity discussion

Comparing to existing works, FAN-Trans is less complex and it has far fewer parameters (learnable or otherwise). In terms of the network structure, SRERL [3] contains VGG19, a cropping module and GGNN; TransAU [2] uses InceptionV3, ROI attention module and transformer, and JÂANet [4] combines hierarchical and multi-scale region learning, face alignment and global feature learning and adaptive attention learning.

Model size: Most works listed in Table 5 did not disclose their model size. To the best of our knowledge, InceptionV3 used in TransAU [2] has $\sim 27M$ parameters, and JÂANet has $\sim 25M$ parameters. In comparison, FAN-Trans has only $\sim 12.15M$ frozen parameters (from FAN), plus $\sim 4.5M$ learnable parameters during training or $\sim 2.8M$ parameters at interference time.

Flops: Except for the face alignment module (13.97G flops), the proposed module only consumes 3.03G flops, which is lower than 8.38G flops required by JAÂNet.

Training Ttime: For training one subset of BP4D, BL takes 65 mins while FANTrans takes 73 mins.

4. More ablation studies

To show the effectiveness of proposed learnable attention drop, we compare it with full attention in the no knowledge distillation loss setting. In practice, we compare them under One-to-One classifier and One-to-Many classifier with a hybrid network to learn AU features. The quantitative results in Table 8 show that in both cases, the learnable attention mechanism is superior to its counterpart, and it obtains 0.8% average F1-score improvement for One-to-One classifier and 0.6% for One-to-Many classifier.

Methods	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU14	AU15	AU17	AU23	AU24	AVG
C_T_C_{o2o}_F	49.5	42.9	59.7	76.6	73.2	79.7	87.1	58.5	50.6	64.4	49.8	52.3	62.0
C_T_C_{o2o}	55.8	44.9	56.9	77.8	75.6	82.8	87.5	61.3	48.7	61.9	48.2	52.5	62.8
C_T_C_{o2m}_F	52.1	48.6	58.9	75.7	73.1	82.9	87.0	61.7	47.6	63.3	46.5	55.2	62.7
C_T_C_{o2m}	52.7	47.0	56.8	75.0	75.3	82.2	88.0	63.1	51.9	64.3	49.5	53.7	63.3

Table 8. Ablation Studies on BP4D. The performance of hybrid of convolution and transformer layers with full precision attention map.

References

[1] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *ICCV*, 2017.

[2] Geethu Miriam Jacob and Bjorn Stenger. Facial action unit detection with transformers. In *CVPR*, 2021.

[3] Guanbin Li, Xin Zhu, Yirui Zeng, Qing Wang, and Liang Lin. Semantic relationships guided representation learning for facial action unit recognition. In *AAAI*, 2019.

[4] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. Jaanet: joint facial action unit detection and face alignment via adaptive attention. *IJCV*, 2021.