Feature Representation Learning with Adaptive Displacement Generation and Transformer Fusion for Micro-Expression Recognition — Supplemental Material —

Zhijun Zhai[1], Jianhui Zhao[1]*, Chengjiang Long[2], Wenju Xu[3], Shuangjiang He[4], Huijuan Zhao[4]
[1]School of Computer Science, Wuhan University, Wuhan, Hubei, China
[2]Meta Reality Labs, Burlingame, CA, USA
[3]OPPO US Research Center, InnoPeak Technology Inc, Palo Alto, CA, USA
[4]FiberHome Telecommunication Technologies Co., Ltd, Wuhan, Hubei, China
zhijunzhai@whu.edu.cn, jianhuizhao@whu.edu.cn, clong1@meta.com, wenjuxu123@gmail.com

Abstract

In this supplementary material, we provide more details of the network configuration, visualization images and experimental results that could not be included in the main article due to the space limitation.

1. Displacement Generation Module

Detailed configuration. The detailed configuration and parameters of each layer of our Displacement Generation Module (DGM) are shown in Table 1. The layers of Conv1-9 are followed by Batch Normalization (BN) layer and Leaky Relu (LR) activation function layer, while the last Conv10 layer is followed by Tanh layers. Up stands for the Upsampling layer using bilinear interpolation. In Concat layer, the output of Conv3 is concatenated with the output of Conv7 and fed into the next layer.

Normalization. In order to normalize the displacements with different scales, we divide each displacement by the average of the first k large values of all its absolute values. The reason for taking the average instead of the maximum value is to reduce the impact of possible large noises. We also include a comparison with a threshold of $\epsilon = 0.0001$ to prevent the division-by-zero error. The normalized image I_{norm} can be obtained from:

$$I_{norm} = \frac{I_{raw}}{max(Avg(|I_{raw}|, k), \epsilon)}, \tag{1}$$

where I_{raw} is the displacement, and function $Avg(I, k)$ represents the average of the first $k = 5$ large values of array I. The normalized displacement is concatenated with the onset-apex frame pair and used as input to the following Transformer Fusion module, preserving both temporal and spatial features.

Layer	Filter	Stride	Output
Input	-	-	224×224 × 2
Conv1	3×3	1	224×224 × 32
BN & LR	-	-	224×224 × 32
Conv2	4×4	2	112×112 × 64
BN & LR	-	-	112×112 × 64
Conv3	3×3	1	112×112 × 64
BN & LR	-	-	112×112 × 64
Conv4	4×4	2	56×56 × 128
BN & LR	-	-	56×56 × 128
Conv5	3×3	1	56×56 × 128
BN & LR	-	-	56×56 × 128
Conv6	3×3	1	56×56 × 128
BN & LR	-	-	56×56 × 128
Up	-	-	112×112 × 128
Conv7	1×1	1	112×112 × 64
BN & LR	-	-	112×112 × 64
Concat	-	-	112×112 × 128
Conv8	3×3	1	112×112 × 64
BN & LR	-	-	112×112 × 64
Conv9	3×3	1	112×112 × 64
BN & LR	-	-	112×112 × 64
Up	-	-	224×224 × 64
Conv10	3×3	1	224×224 × 2
Tanh	-	-	224×224 × 2

Table 1. Detailed configuration of DGM.

Randomization. During training process, we also perform a random perturbation of the apex frame index to augment training data. The selected apex frame index I_{select} for each step is calculated by the following formulas:

$$I_{low} = max(I_i + 1, I_j - \lfloor (I_j - I_i) \times r \rfloor), \tag{2}$$

$$I_{high} = min(I_j + \lfloor (I_j - I_i) \times r \rfloor, I_{last}), \tag{3}$$

$$I_{select} = Random[I_{low}, I_{high}], \tag{4}$$

*Corresponding author.

where $r = 0.4$ is the scale of the optional range, I_i is the onset frame index, I_j is the apex frame index, I_{low} is the lower bound of a random index, I_{high} is its upper bound, I_{last} is the index of the last frame, and $Random$ represents the random selection of an integer value within the range.

2. Transformer Fusion

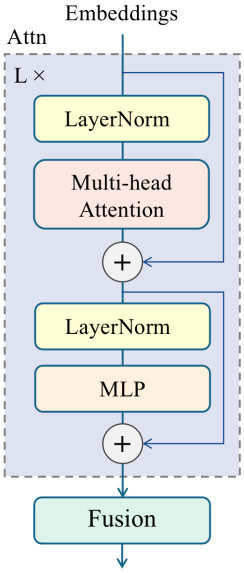


Figure 1. Structure of the Fusion Module.

We provide the structure of the Fusion Modules in our Transformer Fusion in Figure 1. The Attention Learning (Attn) block is a stack of multiple Transformer layers, including Layer Normalization, Multi-head Attention and Multi-layer Perceptron with skip connections, and the output vectors are fed to the subsequent fusion layer for integration. Specifically, we set $L = 2$ for the Local and Global Fusion Modules, and $L = 4$ for the Full-face Fusion Module.

3. AU Regions

To clarify, we design our AU regions based on psychologists Ekman and Friesen's Facial Action Coding System (FACS) [2] [1] which describes the correlation between facial expressions and our AU Regions, *i.e.*,

- **Negative**: eyebrows drooping, eyebrows tightened, squinting, sniffing, corners of mouth pulled down, chin raised or tightened, represented by AU Regions #1, #2, #3, #4, #5 and #6;
- **Positive**: eyelids contracted, tail of eyes formed with crow's feet, cheeks raised and wrinkled, corners of mouth cocked up, represented by AU Regions #1, #3 and #5;
- **Surprise**: wide eyes, eyelids, and eyebrows slightly raised, lips and mouth relaxed, jaw drooping, represented by AU Regions #1, #2, #5 and #6.

In our experiments, for efficiency, we pre-calculate the center positions of each AU region on all sample images in the datasets and take the average as the center of the corresponding AU bounding box used for training.

4. Supplements for Visualization

We add more visualization results of the fusion weights of our local module, global module and full-face module are visualized in Figure 2.

5. Supplements for Failure Cases

In all the failure cases, the main influencing factors are wearing glasses, blinking and eye movements. The classification accuracy of all samples with the three factors is 0.878, while the accuracy of other samples is 0.894. Some of the failure cases influenced by these factors are shown in Figure 3 (a).

Additionally, we find some indistinguishable samples in SMIC dataset whose ground truth labels do not match the intuitive expression category judgments, as shown in Figure 3 (b). These samples may confuse the classification of our network and affect the overall accuracy.

References

[1] Paul Ekman. Facial expression and emotion. *American psychologist*, 48(4):384, 1993. 2

[2] Paul Ekman and Wallace V Friesen. Facial action coding system (facs): A technique for the measurement of facial action. *Rivista Di Psichiatria*, 47(2):126–138, 1978. 2

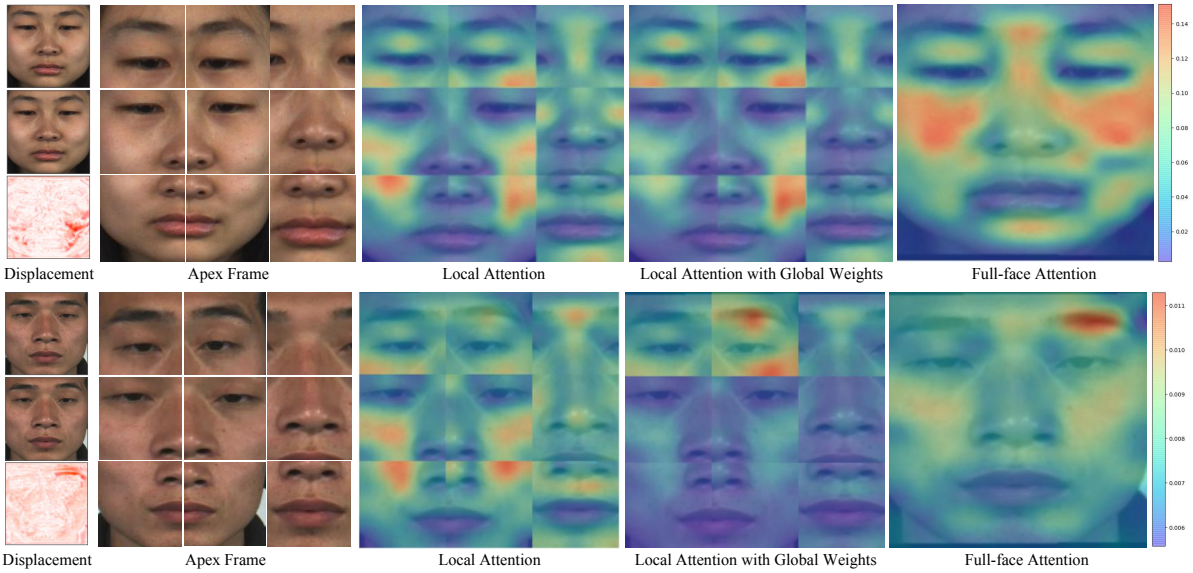
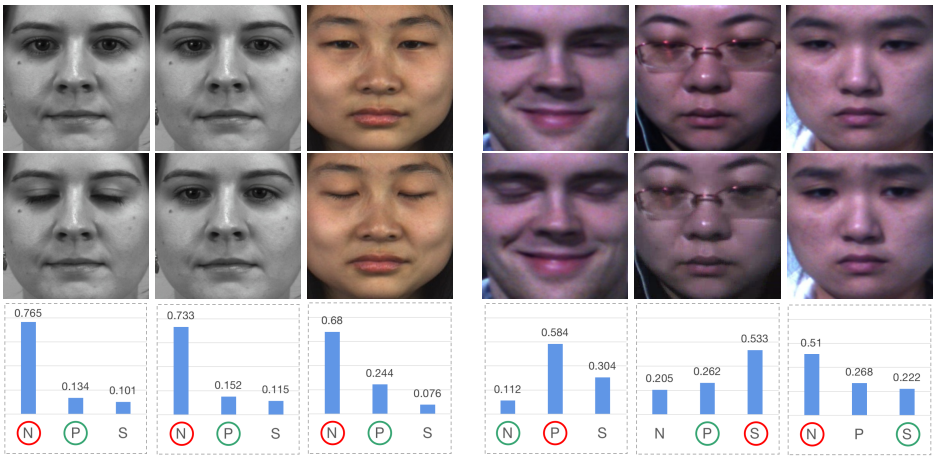


Figure 2. Visualization of the weights in fusion layers. (Images from CASME II ©Xiaolan Fu). Note that the author Zhijun Zhai produced the experimental results in this supplemental material. Meta did not have access to these data.



(a) (b)

Figure 3. Failure cases of our FRL-DGT. Green circle represents for the ground truth. (Images from SAMM, SMIC, and CASME II ©Xiaolan Fu). Note that the author Zhijun Zhai produced the experimental results in this supplemental material. Meta did not have access to these data.