POSTER: A Pyramid Cross-Fusion Transformer Network for Facial Expression Recognition —*Supplementary Material*

Ce Zheng, Matias Mendieta, Chen Chen
Center for Research in Computer Vision, University of Central Florida
{ce.zheng,matias.mendieta}@ucf.edu, chen.chen@crcv.ucf.edu

1. Overview

The supplementary material contains more ablation studies, which are organized into the following sections:

- Section 2: Pyramid Layers
- Section 3: Cross-fusion Mechanism
- Section 4: Model Size and FLOPs Comparison
- Section 5: Transformer encoders depth.
- Section 6: Confusion Matrices

2. Pyramid Layers:

To investigate the optimal pyramid layers with embedded dimensions, we conduct the experiments on the RAF-DB dataset and the results are shown in Table 1. The three pyramid layers with embedded dimensions [512, 256, 128] achieve similar results to the four pyramid layers with embedded dimensions [512, 256, 128]. Considering the computational budget, POSTER adopts the three pyramid layers with embedded dimensions [512, 256, 128] as the optimal choice.

Table 1: Ablation study on the Pyramid Layers.

	RAF-DB	
layers	Acc	mAcc
[512]	91.63	85.01
[512, 256]	91.77	85.49
[512, 256, 128]	92.05	86.03
[512, 256, 128, 64]	92.04	85.97

3. Cross-fusion Mechanism

Table 2: Ablation study on Cross-fusion Mechanism.

	RAF-DB	
	Acc	mAcc
No swap	91.27	85.66
swap for the first block	91.68	85.71
swap for the first two blocks	91.89	85.88
swap for the first four blocks	91.91	85.86
swap for all 8 blocks	92.05	86.03

For POSTER, we swap Q_{img} and Q_{lm} for cross fusion during transformer attention for each MSA block. Image features can be guided by some prior knowledge of salient regions from the landmarks. Likewise, the landmark features are provided with additional global context from the image features. In this way, we foster improved contextual understanding to alleviate intra-class discrepancy and inter-class similarity. We also conduct experiments to evaluate different cross-fusion mechanisms in Table 2. Based on the results, swapping Q_{img} and Q_{lm} for all blocks achieves the best performance.

4. Model Size and FLOPs Comparison:

Previous methods did not pay much attention to the model's computational and memory complexity. The total number of parameters (Params) and floating-point operations (FLOPs) of the model are two key characteristics for a fair comparison, but are often neglected. Furthermore, many recent papers did not release their implementation code. Here we list the Params and FLOPs of DMUE [3] (estimated based on their released code) and TransFER [4] (provided by the author) compared with our POSTER in Table 3. We introduce three versions of our POSTER: POSTER-T (tiny version, the depth of transformer encoders is 4), POSTER-S (small version, the depth of transformer encoders is 6), and POSTER (the depth of transformer encoders is 8). The *Params and FLOPs of the image backbone and landmark detector are included for our methods.*

POSTER-T has lower Params and similar FLOPs compared with DMUE [3], but POSTER-T has much better performance both on RAF-DB and AffectNet datasets. When comparing with TransFER [4], POSTER-T outperforms

Table 3: Comparison on Parameters and FLOPs. The image backbone (IR50) and facial landmark detector (MobileFaceNet) are taken into account when computing Params and FLOPs of POSTER-T, POSTER-S, and POSTER.

Methods	Year	Params	FLOPs	Acc(RAF-DB)	Acc(AffectNet)
DMUE[3]	CVPR 2021	78.4M	13.4G	89.42	63.11
TransFER[4]	ICCV 2021	65.2M	15.3G	90.91	66.23
POSTER-T	-	52.2M	13.6G	91.36	66.86
POSTER-S	-	62.0M	14.7G	91.54	67.13
POSTER	-	71.8M	15.7G	92.05	67.31

Table 4: The trade-off between the complexity versus the performance of POSTER.

			transformer blocks		overall (with backbones)		RAF-DB	AffectNet_7cls
	# of blocks	emb_dim	Params(M)	FLOPs (G)	Params(M)	FLOPs (G)	Acc	Acc
Image only	8 (single ViT)	512	19.7	2.4	50.8	13.1	90.51	65.35
POSTER-T	4 (Cross-Fusion)	512	19.7	2.4	52.2	13.6	91.36	66.86
POSTER-S	6 (Cross-Fusion)	512	29.5	3.5	62.0	14.7	91.54	67.13
POSTER	8 (Cross-Fusion)	512	39.3	4.5	71.8	15.7	92.05	67.31

TransFER for all aspects. If the goal is to pursue higher performance, POSTER would be a good choice since computational and memory complexity is similar to other methods while achieving higher accuracy.

To investigate the trade-off between complexity versus performance, we show the complexity metrics in Table 4. When only using image features modeled by conventional transformer encoders, the Accuracy is 90.51 % on RAF-DB and 65.35 % on AffectNet_7cls. Within the same computational budget of transformer blocks and a similar overall computational budget, POSTER-T outperforms image_only case on both RAF-DB and AffectNet datasets. POSTER-T, POSTER-S, and POSTER have identical image backbone and landmark detector. The only difference between POSTER-T, POSTER-S, and POSTER is the number of transformer blocks. POSTER achieves the best results with 8 cross-fusion transformer blocks.

5. Transformer encoders depth:

In Fig. 1, we plot relations between the accuracy with the network depth. When the number of transformer encoders is greater than 4, the performance is at a relatively high level on both RAF-DB and AffectNet datasets. We choose the depth number to be 8 in our final architecture since this provides the best results.

6. Confusion Matrices:

We show the confusion matrices of the baseline method and the proposed POSTER on RAF-DB, AffectNet (7 cls), and AffectNet (8 cls) datasets in Fig. 2 (a) and (b).

Compared with the baseline, POSTER significantly improves the class-wise accuracy (diagonals of each confusion matrix) on all three experiments in Fig. 2 which indicates that POSTER reduces intra-class discrepancy for FER. Given the target categories, the error rate of predicting into wrong categories also decreases most of the cases when comparing the same positions (except diagonals of each confusion matrix) between Fig. 2 (a) and (b), which shows that POSTER alleviates inter-class similarity for FER.

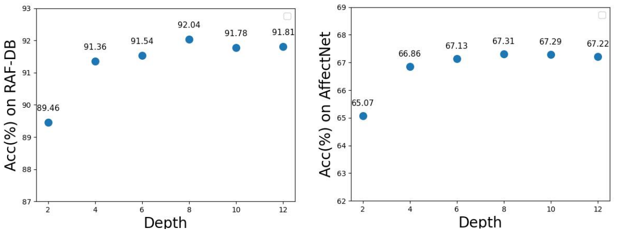
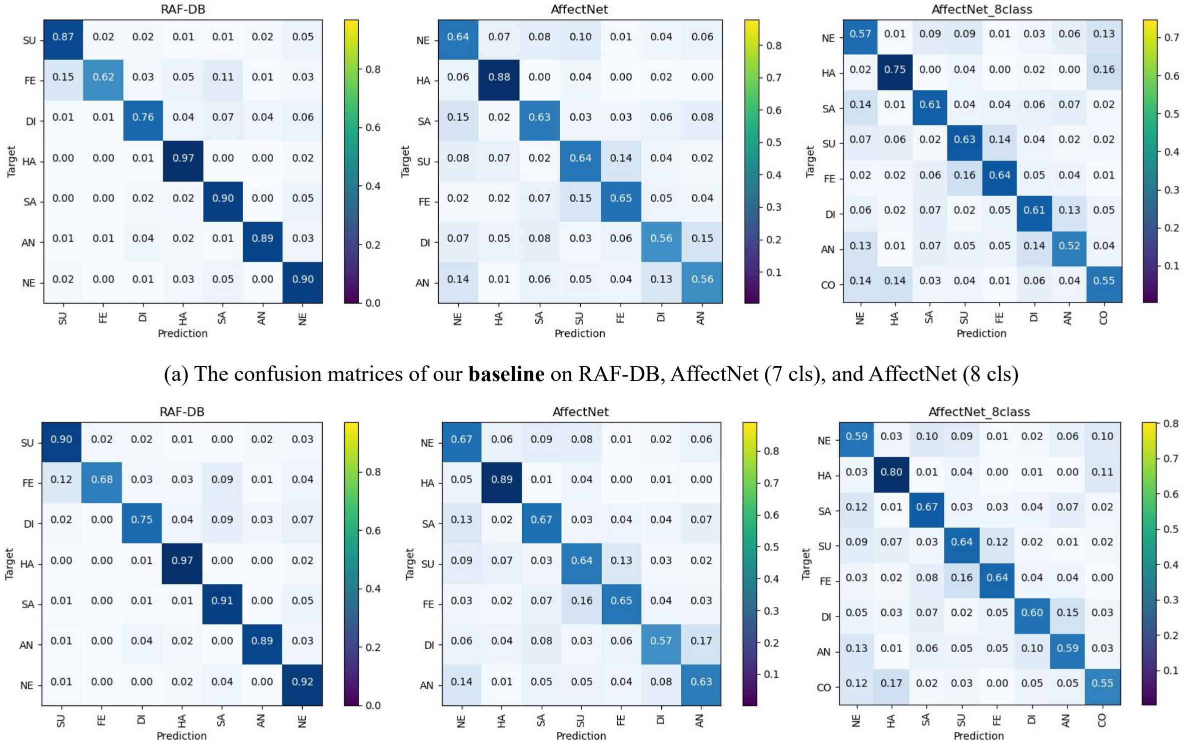


Figure 1: Evaluation of different numbers of transformer encoders (depth) on RAF-DB and AffectNet (7 cls) datasets.



(a) The confusion matrices of our **baseline** on RAF-DB, AffectNet (7 cls), and AffectNet (8 cls)

(b) The confusion matrices of **POSTER** on RAF-DB, AffectNet (7 cls), and AffectNet (8 cls)

Figure 2: Confusion matrices of our baseline (a) and Poster (b) on RAF-DB [1], AffectNet-7cls [2], and AffectNet-8cls [2] datasets

References

[1] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2584–2593. IEEE, 2017. 3

[2] 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. 3

[3] Jiahui She, Yibo Hu, Hailin Shi, Jun Wang, Qiu Shen, and Tao Mei. Dive into ambiguity: latent distribution mining and pairwise uncertainty estimation for facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6248–6257, 2021. 1, 2

[4] Fanglei Xue, Qiangchang Wang, and Guodong Guo. Transfer: Learning relation-aware facial expression representations with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3601–3610, October 2021. 1, 2