Generalizable Facial Expression Recognition Supplementary Material

Yuhang Zhang, Xiuqi Zheng, Chenyi Liang, Jiani Hu, and Weihong Deng

Beijing University of Posts and Telecommunications
{zyhzyh, xiuqizheng, liangchenyi, jnhu, whdeng}@bupt.edu.cn

1 Pipeline of Our Method

We summarize the pipeline of our method in Alg. 1. We utilize fixed pre-trained large model f_{fixed} to extract generalizable face features for any given facial expression recognition (FER) samples. To mimic how humans recognize expressions, we train a newly initialized FER model f from scratch to extract useful expression-related features from the given face features. We assume that if the FER model learns how to extract expression features from given face features, it can generalize to other unseen FER domains based on the generalizable face features extracted from pre-trained large models. We input the learned mask of the FER model to a sigmoid function to reduce overfitting. The learned sigmoid mask is multiplied with the fixed face features to select useful expression-related features. To make the selected features generalizable, we separate them according to the channel dimension into seven pieces corresponding to the seven basic expression classes. Each piece of the selected features represents the features of a certain expression class. To further increase the generalization ability, we make the channels within each piece as diverse as possible through a diverse loss. The training loss is calculated by summing the classification loss, separation loss, and the diverse loss.

2 Implementation Details on AffectNet

Implementation details are slightly different on AffectNet dataset as it is large-scale and imbalanced. The learning rate is 0.0001 and the gamma of the scheduler is 0.8. As AffectNet is extremely imbalanced, we adopt a balanced sampler to keep the samples of each class similar within each batch. The training epoch is 20 instead of 60 because AffectNet has much more training samples than RAF-DB.

3 Mean Accuracy

We display the accuracy of each class on FERPlus in Table 1. From the results, we conclude that our method achieves the best accuracy on most of the expression classes. Our method also performs the best regards of the mean accuracy of all classes.

Algorithm 1 Training Algorithm

Require: Fixed large model f_{fixed}, FER dataset D, newly initialized FER model f with parameters Θ, learning rate η, total epochs $T_{\max}$, data loader iterations I_{train}.

for $t = 1$ to $T_{\max}$ **do**
  for $n = 1$ to I_{train} **do**
    Fetch mini-batch D_n from D
    Extract fixed face features $\mathbf{F}$ using f_{fixed}
    Extract facial expression features $\mathbf{f}$ using f
    Calculate sigmoid mask $\mathbf{M}$ by resizing $\mathbf{f}$ and inputting it to the sigmoid function
    Obtain the masked feature $\tilde{\mathbf{F}}$ using Eq. (2)
    Calculate the classification loss ℓ_{cls} using Eq. (3)
    Separate the masked feature $\tilde{\mathbf{F}}$ and apply channel dropping as per Eq. (4)
    Calculate the logits without FC layer using Eq. (5)
    Calculate the separation loss ℓ_{sep} using Eq. (6)
    Separate and max pool the masked feature $\tilde{\mathbf{F}}$ according to Eq. (7)
    Calculate the diverse loss ℓ_{div} using Eq. (8)
    Calculate the training loss ℓ_{train} using Eq. (9)
    Update $\Theta = \Theta - \eta \nabla \ell_{train}$
  end for
end for

Ensure: The trained FER model f, which can selectively extract expression-related features from the given fixed face features.

Table 1: The performance on the FERPlus test set with the accuracy of each expression class, when the train set is RAF-DB. Overall accuracy is the accuracy on the whole test set, mean accuracy is the mean value of the accuracy of each expression class. Our method achieves both the best overall and the mean accuracy compared with other methods.

Method	Surprise	Fear	Disgust	Happiness	Sadness	Anger	Neutral	Overall	Mean
SCN	**82.83**	14.46	27.78	77.72	61.20	**61.17**	35.78	58.37	51.56
RUL	69.70	**45.78**	22.22	82.53	70.57	52.01	31.93	57.89	53.54
EAC	70.71	39.76	27.78	77.94	76.04	58.24	22.11	54.38	53.23
OFER	71.97	36.14	27.78	76.15	73.44	53.85	24.04	53.90	51.91
CAFE	78.54	37.35	**33.33**	**94.96**	**78.65**	57.51	**58.72**	**73.16**	**62.72**

4 Visualizaton of Error Samples

We observe that when trained on RAF-DB, the neutral class of FERPlus is easy to be misclassified, while the fear class of the other three datasets is easy to be misclassified. Neutral and fear classes are both semantically ambiguous. Some examples are shown in Fig. 1 to help understanding.

Fig. 1: Some error examples on unseen datasets. Labels are displayed in black and incorrect predictions in red.

5 Performance Without Pre-Trained FER Model

To further illustrate the effectiveness of our proposed method CAFE, we carry out experiments without using the pre-trained backbone of FER model.

Table 2: Comparison between with or without pre-trained FER backbone. The test accuracy of different FER methods on various FER test sets is shown. We underline the best accuracy of other FER methods and highlight the improvement achieved by our method compared to it in blue. Our method outperforms SOTA FER methods by even larger margins when without a pre-trained FER backbone.

Method	RAF-DB	FERPlus	AffectNet	SFEW2.0	MMA	Mean
With Pre-Trained FER Backbone						
SCN	87.32	58.37	42.85	44.89	36.52	53.99
RUL	88.66	57.89	43.82	46.91	37.11	54.88
EAC	**89.54**	54.38	43.91	43.39	37.27	53.70
OFER	89.07	53.90	42.73	43.88	36.43	53.20
CAFE	88.72	**73.16**	**45.86**	**52.86**	**56.80**	**63.48**
	-0.82	+14.79	+1.95	+5.95	+19.53	+8.60
Without Pre-Trained FER Backbone						
SCN	75.23	38.48	21.41	17.17	30.11	36.48
RUL	79.53	35.86	16.42	11.78	11.89	31.10
EAC	80.64	41.73	23.73	22.22	29.51	39.57
OFER	80.17	35.66	20.83	18.90	28.15	36.74
CAFE	**85.92**	**69.14**	**40.96**	**45.79**	**54.65**	**59.29**
	+5.28	+27.41	+17.23	+23.57	+24.54	+19.72

All the FER methods are trained from scratch without a pre-trained backbone on RAF-DB and tested on all five different FER test sets. The results are shown in Table 2. We observe that our method outperforms the SOTA FER methods on the generalization ability by even larger margins when without a pre-trained FER backbone. For example, without a pre-trained FER backbone, our model increases the second-best performance on FERPlus by 27.41% compared with 14.79% when with the pre-trained FER backbone. The performance of all FER methods drops when without the pre-trained FER backbone, however, our method achieves a very similar performance between the two groups, only decreasing the mean performance from 63.48% to 59.29%, which is acceptable. While the EAC method decreases the mean performance from 53.70% to 39.57%, which is rather drastic. The results illustrate that the pre-trained FER backbone is not necessary for our method to achieve the best performance. Our method without a pre-trained FER backbone still achieves a mean accuracy of 59.29%, outperforming the best mean accuracy of other methods with a pre-trained FER backbone of 54.88%. From the comparison between the two groups, we conclude that our method is robust and performs well even without the pre-trained FER backbone. The reason lies in that our method only learns a mask to select fixed face features instead of learning the whole expression features, which does not need a very strong FER model to perform well. Furthermore, a simple FER model without pre-training can extract simple patterns to select fixed face features, which is more likely to generalize across different FER datasets.