Generalizable Facial Expression Recognition

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

Abstract. SOTA facial expression recognition (FER) methods fail on test sets that have domain gaps with the train set. Recent domain adaptation FER methods need to acquire labeled or unlabeled samples of target domains to fine-tune the FER model, which might be infeasible in real-world deployment. *In this paper, we aim to improve the zero-shot generalization ability of FER methods on different unseen test sets using* ***only one train set****.* Inspired by how humans first detect faces and then select expression features, we propose a novel FER pipeline to extract expression-related features from any given face images. Our method is based on the generalizable face features extracted by large models like CLIP. However, it is non-trivial to adapt the general features of CLIP for specific tasks like FER. To preserve the generalization ability of CLIP and the high precision of the FER model, we design a novel approach that learns sigmoid masks based on the fixed CLIP face features to extract expression features. To further improve the generalization ability on unseen test sets, we separate the channels of the learned masked features according to the expression classes to directly generate logits and avoid using the FC layer to reduce overfitting. We also introduce a channel-diverse loss to make the learned masks separated. Extensive experiments on five different FER datasets verify that our method outperforms SOTA FER methods by large margins. Code is available in https://github.com/zyh-uaiaaaa/Generalizable-FER.

Keywords: Facial expression recognition · Zero-shot generalization · Mask learning

1 Introduction

Facial expression recognition (FER) aims to understand human feelings and is vital to human-computer interaction. With the development of deep learning, FER methods achieve great success on both laboratory-collected and in-the-wild FER datasets. However, we find that things are different when the test sets have domain gaps with the train set. For example, if we train the SOTA FER model EAC [49] on one given FER dataset like RAF-DB [25], then it can only achieve high performance on the test set of RAF-DB, while very low performance on other different FER test sets like AffectNet [29], shown in Fig. 1. Though some works try to deal with the domain generalization problem in FER, they all assume accessing labeled or unlabeled target set samples for FER model
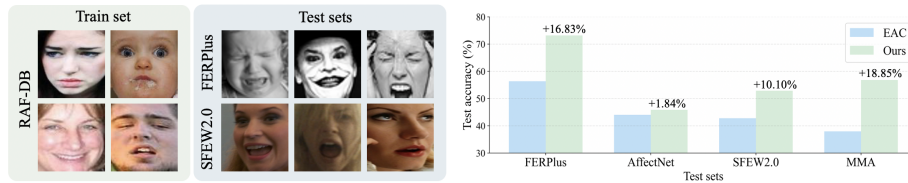
Fig. 1: The illustration of Generalizable Facial Expression Recognition (GFER). In order to evaluate the generalization ability of FER methods, we train FER models on one train set and test the trained models on different unseen test sets. *The difference between our task and domain adaptation FER is that we only use* **one** *train set and* **do not** *acquire any labeled or unlabeled samples from the target domain.* SOTA FER methods like EAC do not work well on unseen test sets, which show low generalization ability. Our method outperforms EAC by large margins on different unseen FER test sets, showing better generalization ability.

fine-tuning. However, in the real-world FER, as we do not previously know the distribution of the test samples, we might not be able to get access to even the unlabeled target samples. In such cases, the domain adaptation FER methods cannot work. *In this paper, we aim to improve the zero-shot generalization ability of FER methods on different unseen test sets using only one train set.* The difference between our work and existing domain adaptation FER works is that we only utilize **one** given train set and we **do not** utilized labeled or unlabeled target samples to fine-tune our model. This is more similar to the real-world deployment where we usually only have one train set with a similar style of annotating (*without considering annotation inconsistency*) and we need to recognize unseen test samples from various domains. As the generalization ability of the FER model is our main focus, we name the paper as Generalizable Facial Expression Recognition (GFER).

We find in our experiments that SOTA FER methods achieve low test accuracy on different unseen FER test sets. The reason might lie in that these methods fit the given FER train set too well (including domain-specific information) to predict on test sets with domain gaps. However, we humans take a very generalizable way to recognize facial expressions. Given images with domain gaps, we first locate the face features, then we disentangle expression features from the face features and make a recognition only based on the expression features. Inspired by that, we design a novel pipeline to improve the generalization ability of FER methods. We first extract generalizable face features from pre-trained large models like CLIP. As the large models have been trained with numerous images including face images, we can assume the extracted face features from them could generalize seamlessly to different FER domains. However, it is non-trivial to adapt the general features extracted by CLIP for specific tasks like FER. We further propose a novel method to learn sigmoid masks to select expression-related features for FER. Specifically, we fix the general face features during the whole training process to maintain the generalization ability. We then

train a FER model to learn masks based on these general face features. In order to make the learned masks generalizable instead of overfitting the given face features, we propose to utilize a sigmoid function to regularize the learned masks. The sigmoid mask is vital to the success of method, which is discussed in our experiments. Furthermore, enlightened by the channel-wise attention and the fixed number of the seven basic expressions, we separate the masks to respond to different expressions according to the channels. This operation avoids using the FC layer, which further reduces the overfitting probability of the FER model. Finally, we introduce a channel-diverse loss to make the masks corresponding to different expressions as diverse as possible to improve the generalization ability of the learned masks.

Extensive experiments on five different FER datasets validate the effectiveness of our proposed method. We carry out ablation study and design comparison groups to study each part of our method. Visualization results of learned features and sigmoid masks are provided to help understanding. We summarize our main contributions as follows.

- We propose to learn sigmoid masks on fixed general face features for FER. This novel FER pipeline improves zero-shot generalization ability on different unseen test sets using only one train set.
- We design a channel-separation module and a channel-diverse loss to further improve the generalization ability of the learned sigmoid masks. Channel-separation modules separate the learned masked features according to the channel dimension to reduce overfitting. The channel-diverse loss regularizes the sigmoid masks to be as diverse as possible.
- Extensive experiments on five different FER methods illustrate that our method outperforms SOTA FER methods by remarkable margins on different unseen test sets, which shows the superior generalization ability of our proposed method.

2 Related Work

2.1 Facial Expression Recognition

Facial Expression Recognition (FER) plays a vital role in human-computer interaction, and extensive research has been conducted to enhance the performance of FER [1, 7, 14, 15, 17, 21–25, 34, 36, 47, 50, 51]. For instance, Li *et al.* [25] use crowd-sourcing to simulate human expression recognition, while [1, 15] employ model ensembling to leverage more information. Farzaneh *et al.* [7] propose a center loss variant to maximize intra-class similarity and inter-class separation for FER, and Ruan *et al.* [34] acquire expression-relevant information during the decomposition of an expression feature. However, we find that these FER methods are effective when the test set has no domain gap with the train set, while their performance drops drastically when facing domain-different test sets. In this paper, we aim to improve the generalization ability of the FER model and make it suitable for real-world deployment. We train the FER model on one given FER train set and evaluate it on all unseen different FER test sets.

2.2   Domain Generalization

Domain Generalization (DG) aims to help models trained on a set of source domains generalize better on unseen target domains [5, 8, 12, 16, 20, 27, 28, 33, 39, 40, 44–46]. A common practice is to reduce the feature discrepancy among multiple source domains. [26, 28, 41] all adopt maximum mean discrepancy on multiple layers to enforce the distribution similarity between source and target features. Deep CORAL [38] uses feature covariance to measure the domain discrepancy. Another stream of works tries to enlarge the available train data space with augmentation of source domains [3, 6, 30, 37, 52, 53]. Several approaches leverage regularization through domain adversarial learning [13, 32] to address DG. However, domain generalization (DG) methods in the FER field usually assume the availability of labeled or unlabeled target expression samples to aid in fine-tuning the FER models. However, our approach differs significantly. We strive to enhance the generalization ability of facial expression recognition (FER) methods by exclusively training the FER model on a single FER dataset and evaluating it on various unseen FER test sets. We refrain from accessing any samples from the target domain for fine-tuning our method, rendering existing FER domain adaptation methods infeasible for our task.

3   Problem Definition

In this paper, we aim to improve the generalization ability of the FER model and guide it to recognize unseen test samples with domain gaps of the train samples. Learning is conducted on one given FER train set, and then test samples from different FER test sets with domain gaps of the train set should be recognized on the fly, which is similar to the real-world deployment of FER models.

FER models are trained with $\mathcal{D}_{train} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, where $\mathbf{x}_i$ is the i-th training image and $y_i \in Y = \{1, \ldots, L\}$ is the corresponding label, N is the number of training samples and L is the number of expression classes, we consider the seven basic expressions across different FER datasets in this paper, thus L equals 7. Existing FER models are evaluated on the test set $\mathcal{D}_{test} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{M}$ that has no domain gap between the train set, M is the number of test samples. However, the real-world test set $\mathcal{D}_{real} = \{(\widetilde{\mathbf{x}}_i, y_i)\}_{i=1}^{M}$ is different from $\mathcal{D}_{test}$, as $\mathcal{D}_{real}$ might contain samples with domain gaps of the training samples. In this paper, we aim to train the FER model on $\mathcal{D}_{train}$ that can generalize well on the real-world test set $\mathcal{D}_{real}$. The difference between our setting and the traditional FER is that we test the FER models on $\mathcal{D}_{real}$ instead of $\mathcal{D}_{test}$. The difference between our setting and domain adaptation FER is that we only use the training set from one source domain and do not have access to the unlabeled samples from $\mathcal{D}_{real}$ to fine-tune the FER model. Our setting is more similar to the real-world deployment as we cannot previously know the distribution of the target domain in the real world.

4 Method

To solve the aforementioned problem, we propose a novel method to mimic the <u>C</u>ognition of hum<u>A</u>n for <u>F</u>acial <u>E</u>xpression and name it as CAFE. As humans, when we encounter test samples with domain gaps from the training samples, we first extract their face features to exclude the domain features. Afterward, we focus solely on the features related to the expressions in order to determine the expression contained within the test sample. Following a similar approach, the proposed method initially extracts the face features of the test samples. Then, a trained FER model generates masks for selecting expression-related features from the facial features and making decisions solely based on the selected features. The framework of our proposed method is shown in Fig. 2.

4.1 Mask on Fixed Face Features

We design a novel method to guide the model to learn masks on fixed face features in order to selectively choose useful expression features. Since face features can be extracted using pre-trained large models like CLIP [31], we can assume that these features are sufficiently generalizable for different domains.

Given images $\mathbf{x}$ from $\mathcal{D}_{train}$, we first extract the face features using CLIP, denoted as $\mathbf{F} \in \mathbb{R}^{N \times C}$, where N is the number of images and C the number of feature dimensions, we fix $\mathbf{F}$ during the training process to prevent the FER model to optimize face features to overfit the given train set. This operation improves the generalization ability of our proposed method. The FER model, such as ResNet-18, is trained to learn masks for the given face features. We first extract features $\mathbf{f} \in \mathbb{R}^{N \times C \times 1 \times 1}$ from $\mathbf{x}$ after the global average pooling (GAP) layer and resize them to generate the masks $\mathbf{M} \in \mathbb{R}^{N \times C}$ for face features. Further, in order to regularize the learned masks to generalize to unseen test samples, we apply a sigmoid function on $\mathbf{M}$ and get $\mathbf{M}_s$ as

$$\mathbf{M}_s = Sigmoid(\mathbf{M}). \tag{1}$$

The sigmoid function is vital to the success of our method as it introduces non-linearity into the model, which is crucial for capturing and learning non-linear patterns to generate the masks. The sigmoid function also normalizes $\mathbf{M}$, ensuring that the values of $\mathbf{M}$ fall within $[0, 1]$, which reduces the overfitting ability of the learned masks. Furthermore, the sigmoid function provides a probability-like output, where the output value represents the probability of selecting the feature of the corresponding channel, which is semantically similar to humans choosing expression-related features from the face features.

We further utilize $\mathbf{M}_s$ to select the face features for expression recognition as

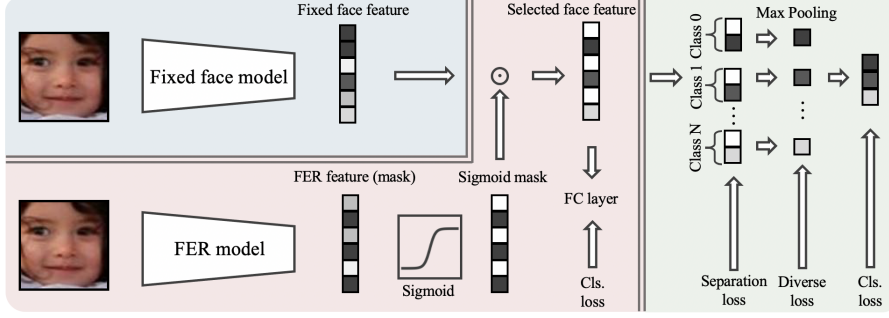$$\widetilde{\mathbf{F}} = \mathbf{M}_s \mathbf{F}. \tag{2}$$

Fig. 2: The framework of our proposed method CAFE. We utilize a fixed pre-trained large model, such as CLIP to extract fixed face features regards of the input training images, the FER model is trained to learn a mask for the fixed face features to only extract the expression-related features. Notice that this is similar to how human perceives expressions: we first observe faces and then extract expression-relevant features. The learned mask is regularized by a sigmoid function to prevent overfitting. We further introduce channel-separation and channel-diverse to make the learned mask diverse to improve the generalization ability.

The classification loss is computed between the selected features $\widetilde{\mathbf{F}}$ and labels y following

$$l_{cls} = -\frac{1}{N} \sum_{i=1}^{N} (\log \frac{e^{\mathbf{W}_{\mathbf{y}_i} \widetilde{\mathbf{F}}_i}}{\sum_j^L e^{\mathbf{W}_j \widetilde{\mathbf{F}}_i}}), \tag{3}$$

where $\widetilde{\mathbf{F}}_i$ is the selected features of image $\mathbf{x}_i$, $\mathbf{W}_{\mathbf{y}_i}$ is the $\mathbf{y}_i$-th weight from the FC layer and $\mathbf{y}_i$ is the label of $\mathbf{x}_i$.

4.2   Mask Generalization

To make the learned mask generalizable to other unseen FER test sets, we further design a separation module to regularize the learned mask. Specifically, we set apart the masked features into seven pieces according to the channel dimension to make the masked features correspond to seven basic expressions and then max pool each piece to directly transform them into logits. In such a design, we could avoid the use of the FC layer and directly connect the masked features to the FER labels. The motivation of our design is three folds: Firstly, the learning ability of the FC layer might be too strong to overfit the train set. Thus, we directly transform the masked features into logits to prevent the FC layer from overfitting the labels with the learned features. Secondly, the channel size of 512 when using ResNet-18 might be too large to learn generalizable masks and could also lead to overfitting on the training set. If we set apart the mask into seven pieces and make each piece of the mask correspond to one basic expression, the mask piece with a small channel size will be more likely to only focus on the useful information. Thirdly, setting the masked features into seven pieces

is similar to label distribution learning, we guide the seven pieces to focus on features related to the seven basic expressions on one image to solve the FER from a label distribution learning perspective. As FER is similar to a label distribution learning task, one image might contain features of several different expressions. For example, there are compound expressions and one compound expression image contains features from several different basic expression classes.

Specifically, given the masked features $\widetilde{\mathbf{F}}$, we divide them according to the channel number C to L pieces corresponding to the class number L on the second dimension, $\widetilde{\mathbf{F}} = \{\widetilde{\mathbf{F}}_1, \widetilde{\mathbf{F}}_2, ..., \widetilde{\mathbf{F}}_L\}$. If C can not be divided by L, we could divide the selected features non-uniformly. For example, when using ResNet-18, the channel size is 512, we could assign 73 channels for each of the 6 expressions and leave the rest 74 channels for the neural expression. We utilize channel-dropping on the selected features to mitigate the overfitting problem of the selected features. The drop mask is denoted as $\mathbf{M}_{drop} = \{\mathbf{M}_1, \mathbf{M}_2, ..., \mathbf{M}_L\}$ with the same size of the selected features. Each mask contains 0 or 1 for feature selection. For example, in the $\mathbf{M}_1$, which is a vector of size $(N, 73)$, N is the batch size, we keep the mask same in each batch, thus, for simplicity, we neglect N in the following. On the channel dimension of 73, if we set the drop rate as $10/73$, there are 10 channels randomly selected as 0 and all the rest are 1. The channel drop module guides the model to focus on all channels, which increases the generalization ability of the masked features. After channel dropping, the selected features are denoted as

$$\overline{\mathbf{F}} = \widetilde{\mathbf{F}}\mathbf{M}_{drop} = \{\widetilde{\mathbf{F}}_1\mathbf{M}_1, \widetilde{\mathbf{F}}_2\mathbf{M}_2, ..., \widetilde{\mathbf{F}}_L\mathbf{M}_L\}. \tag{4}$$

The selected features $\overline{\mathbf{F}}$ is downsized to logits $\overline{\mathbf{F}}^d$ of size (N, L) for classification through a maxpooling operation on the second dimension.

$$\overline{\mathbf{F}}^d = \{max(\widetilde{\mathbf{F}}_1\mathbf{M}_1), max(\widetilde{\mathbf{F}}_2\mathbf{M}_2), ..., max(\widetilde{\mathbf{F}}_L\mathbf{M}_L)\}. \tag{5}$$

Then, we compute a classification loss l_{sep} with the labels and the logits $\overline{\mathbf{F}}^d$ that are obtained by the separation module without the FC layer.

$$l_{sep} = -\frac{1}{N}\sum_{i=1}^{N}(\log \frac{e^{\overline{\mathbf{F}}^d_{\mathbf{y}_i}}}{\sum_j^L e^{\overline{\mathbf{F}}^d_j}}), \tag{6}$$

To increase the generalization ability of the learned masks, we want to make the channels corresponding to each class as diverse as possible. Thus, we further introduce a channel-diverse loss. Specifically, we input the masked features $\widetilde{\mathbf{F}} \in \mathbb{R}^{N \times C}$ into the max pooling operation to get

$$\widetilde{\mathbf{F}}_{max} = \{max(\widetilde{\mathbf{F}}_1), max(\widetilde{\mathbf{F}}_2), ..., max(\widetilde{\mathbf{F}}_L)\}, \tag{7}$$

where $\widetilde{\mathbf{F}}_{max} \in \mathbb{R}^{N \times L}$. The selected max feature channel of the second dimension of $\widetilde{\mathbf{F}}_{max}$ is regularized to be diverse with other feature channels by the channel-diverse loss l_{div},

$$l_{div} = 1 - \frac{1}{Nc}\sum_{i=1}^{N}\sum_{j=1}^{L}\widetilde{\mathbf{F}}_{max_{ij}}, \tag{8}$$

where i, j are utilized to index the first and second dimension of $\widetilde{\mathbf{F}}_{max}$, c is the number utilized for normalization. We experimentally set c as 73, which is the same as the channel number of the separated piece for each expression class. The channel-diverse loss regularizes the max value of each piece of $\widetilde{\mathbf{F}}_{max}$ as large as possible, which separates the largest value of each piece of $\widetilde{\mathbf{F}}_{max}$ from other values, making the values of the channel dimension more diverse.

Combining the separation loss and channel-diverse loss, we aim to learn a powerful while generalizable mask to only select useful expression features from the fixed generalizable face features. The total train loss is summarized as

$$l_{train} = l_{cls} + \lambda l_{sep} + \beta l_{div}. \tag{9}$$

After training, the module to compute l_{sep} and l_{div} can be abandoned, we only need to keep the module that is used to compute l_{cls} to recognize the test samples as l_{sep} and l_{div} are only used to regularize the l_{cls} during training.

5   Experiments

5.1   Datasets

RAF-DB [25] is annotated with seven basic expressions by 40 trained human coders, including 12,271 images for training and 3,068 images for testing.

FERPlus [2] is extended from FER2013 [9] with cleaner labels, which consists of 28,709 training images and 3,589 test images collected by the Google search engine, we utilize the same seven basic expressions with the RAF-DB.

AffectNet [29] is a large-scale FER dataset, which contains eight expressions (seven basic expressions and contempt). There are a total of 286,564 training images and 4,000 test images. We utilize the seven basic expressions in our experiments.

SFEW2.0 is the most commonly used version of SFEW [4]. SFEW2.0 contains 958 train samples, 436 validation samples, and 372 test samples. Each image is assigned to one of the seven basic expressions.

MMA is a large-scale FER dataset with the majority of expressions from individuals of European and American descent. The dataset contains 92,968 training samples, 17,356 validation samples, and 17,356 test samples. Each image is assigned to one of the seven basic expressions.

5.2   Implementation Details

We utilize the most widely employed ResNet-18 [10] as the backbone following other FER works [43,49]. We use the ViT-B/32 CLIP model, only employing its visual component to extract generalizable face features. Any large models trained on extensive face datasets could replace the CLIP model. As studying different models for face feature extraction is not the focus of our paper, we directly utilize the ViT-B/32 CLIP model across all our experiments. The learning rate η is set to 0.0002 and we use Adam [18] optimizer with weight decay of 0.0001.

Table 1: The test accuracy of FER methods on various FER test sets. The model is only trained using the dataset in the second column (source domain) and evaluated on all five test sets. Our method outperforms SOTA FER methods by large margins on unseen test sets (target domains). We underline the best accuracy of all the compared FER methods and bold the best overall performance. Note that we focus on the generalization ability of FER methods instead of the test accuracy on the source domain.

Method	RAF-DB	FERPlus	AffectNet	SFEW2.0	MMA	Mean
SCN (CVPR2020)	87.32	<u>58.37</u>	42.85	44.89	36.52	53.99
RUL (NeurIPS2021)	88.66	57.89	43.82	<u>46.91</u>	37.11	54.88
EAC (ECCV2022)	**89.54**	54.38	<u>43.91</u>	43.39	<u>37.27</u>	53.70
OFER (ICCV2023)	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**

Method	FERPlus	RAF-DB	AffectNet	SFEW2.0	MMA	Mean
SCN (CVPR2020)	86.80	<u>68.71</u>	32.42	43.10	59.12	58.03
RUL (NeurIPS2021)	88.40	51.89	35.88	<u>45.90</u>	58.00	56.01
EAC (ECCV2022)	89.03	58.62	<u>36.49</u>	45.79	<u>59.89</u>	57.96
OFER (ICCV2023)	89.26	55.67	35.84	44.52	59.37	56.93
CAFE	**89.51**	**72.91**	**39.44**	**49.38**	**60.14**	**62.28**

Method	AffectNet	RAF-DB	FERPlus	SFEW2.0	MMA	Mean
SCN (CVPR2020)	62.48	<u>70.70</u>	<u>63.98</u>	41.98	<u>38.51</u>	55.53
RUL (NeurIPS2021)	58.70	55.83	52.88	34.01	31.93	46.67
EAC (ECCV2022)	64.77	66.10	57.19	<u>44.89</u>	33.49	53.09
OFER (ICCV2023)	63.90	63.21	56.79	42.60	31.99	51.70
CAFE	**64.87**	**72.69**	**69.94**	**51.18**	**49.65**	**60.27**

Method	SFEW2.0	RAF-DB	FERPlus	AffectNet	MMA	Mean
SCN (CVPR2020)	51.35	46.74	32.58	23.19	17.38	34.25
RUL (NeurIPS2021)	51.87	46.35	<u>36.02</u>	<u>30.41</u>	22.06	37.34
EAC (ECCV2022)	52.46	<u>47.29</u>	33.66	25.57	<u>22.26</u>	36.25
OFER (ICCV2023)	52.15	46.33	34.78	23.02	21.88	35.63
CAFE	**53.79**	**54.43**	**48.39**	**32.24**	**36.34**	**45.04**

Method	MMA	RAF-DB	FERPlus	SFEW2.0	AffectNet	Mean
SCN (CVPR2020)	63.00	74.09	**<u>73.99</u>**	<u>45.90</u>	<u>35.94</u>	58.58
RUL (NeurIPS2021)	61.70	71.94	69.05	39.21	34.45	55.27
EAC (ECCV2022)	65.06	<u>74.32</u>	71.85	42.87	35.83	57.99
OFER (ICCV2023)	64.74	72.85	69.69	41.49	34.58	56.67
CAFE	**65.97**	**78.36**	73.57	**49.05**	**41.85**	**61.76**

We utilize a learning rate scheduler of ExponentialLR, with a gamma of 0.9. We set the weight for the channel-wise loss as 1.5 and the weight for the diverse loss as 5.

5.3   Main Experiments of GFER

To evaluate the generalization ability of existing FER methods, we utilize one of the five FER datasets as the training set. Instead of testing only on the test set of the corresponding train set, we test on all five FER test sets. Note that we **do not** have access to any labeled or unlabeled images of the target domain. The results are shown in Tab. 1. Our comparison methods are FER methods published in the top conferences in recent years, such as SCN [42], RUL [48], EAC [49], and OFER [19].

Generalization Ability The results reveal that SOTA methods in facial expression recognition (FER) do not exhibit strong performance when applied to FER test datasets characterized by domain gaps relative to the training set. For instance, while an advanced FER method like EAC significantly enhances performance over the SCN method on the test set that corresponds to the training set, its results on other FER test datasets are comparable to, or even worse than, those of the SCN. We speculate the reason might lie in that EAC fits the train set too well, which improves the test accuracy on the corresponding test set and degrades the test accuracy on other different FER test sets. We underline the best result of other FER methods and compare it with our method. Our method outperforms other FER methods by large margins under almost all settings. Under different FER train sets, our method always achieves the best mean accuracy on five different FER test sets. We also show the test accuracy of each expression class and the mean accuracy in the Supp. material and find that our method also achieves the best mean test accuracy of different expression classes.

Cross-Domain Analysis We observe that test accuracy on RAF-DB and FER-Plus are the highest across other datasets. Using SFEW2.0 as the train set achieves the lowest test accuracy on other test sets, as SFEW2.0 is the smallest dataset used in our paper and some images have low quality. Test accuracy on AffectNet is always low given different train sets because the labels of AffectNet are very noisy. Some error examples are shown in the Supp. material to help understanding.

5.4   Discussion

Our paper focuses on the generalization ability of FER methods instead of the test accuracy on the source domain. Actually, we find that strong FER methods like EAC that perform well on the test set of source domain only fits the train set well and cannot generalize to other different unseen test sets. Thus, in this paper, we aim to propose a method that can perform well on the source domain test set and generalize to other unseen test sets at the same time. We do not claim that our method always achieves the best performance on the test set of the source domain, which is not our main focus. Based on the experimental results in Tab. 1, we argue that our method performs on par with SOTA FER methods on different source domain test sets and achieves the best performance on different unseen test sets, which completely fulfills our goal.

Table 2: The ablation study of our proposed method. The results on different FER test sets show that the most effective module of our method is the mask module. The separation and diverse module make the learned mask more generalizable to other FER test sets and further improve the performance of our method.

Mask	Separation	Diverse	FERPlus	AffectNet	SFEW2.0	MMA	Mean
			58.05	43.25	42.76	42.61	46.67
✓			70.90	43.77	51.63	55.65	55.49
✓	✓		72.01	45.17	**53.31**	56.69	56.80
✓	✓	✓	**73.16**	**45.86**	52.86	**56.80**	**57.17**

5.5   Ablation Study

To study the effectiveness of each of the proposed modules in our method, we carry out thorough ablation studies. The results in Tab. 2 show that without our method, the FER model cannot generalize to other datasets that have domain gaps with the train set, which is unsatisfactory for the real-world deployment of FER models. With our proposed sigmoid mask learning, the performance on other unseen test sets outperforms the baseline by a large margin. However, the fitting probability of the mask is too strong as the dimension of 512 might be too much to learn generalizable expression features. Thus, we further introduce the separation module which separates the masked features into pieces corresponding to the number of expression classes. It shows that the results improve from only using the sigmoid mask module. We also introduce a channel-diverse loss to make the channels in each piece as diverse as possible, which further improves the accuracy based on using the sigmoid mask and the channel-separation module. From the results in Tab. 2, each module contributes to the success of our proposed method, and combining them together achieves the best result.

5.6   Different Backbones

We combine our method with different backbones to show its generalization ability. The results in Tab. 3 illustrate that our method improves the performance of baseline under different backbones by remarkable margins. Specifically, the improvement is largest when using ResNet-18 as the backbone, the reason might lie in that the dimension of the output feature of ResNet-18 is 512, which is the same as the dimension of the output feature of CLIP. When using MobileNet [11] or ResNet-50, we reduce the size of the output feature through mean operation to suit the feature dimension of CLIP, which might slightly limit the performance improvement. For example, when the backbone is ResNet-50, we simply reduce the output feature dimension of 2048 to 512 through mean operation using sliding windows. We also observe that stronger backbones have better generalization ability in our experiment. The ResNet-50 backbone achieves the overall best performance across experiments.

Table 3: The performance of our method under different backbones. Our method could improve the performance on unseen FER test sets under different backbones.

Backbone	RAF-DB	FERPlus	AffectNet	SFEW2.0	MMA	Mean
MobileNet	84.65	61.33	**43.45**	43.21	40.67	54.66
MobileNet + CAFE	**85.07**	**64.97**	42.91	**45.45**	**43.11**	**56.30**
ResNet-18	88.40	58.05	43.25	42.76	42.61	55.01
ResNet-18 + CAFE	**88.72**	**73.16**	**45.86**	**52.86**	**56.80**	**63.48**
ResNet-50	88.49	70.13	47.46	49.94	48.90	60.98
ResNet-50 + CAFE	**89.05**	**75.26**	**47.49**	**51.07**	**57.46**	**64.07**

Table 4: Comparison with CLIP+Finetune. The comparison group has the similar trainable parameters with our method for fair comparison. Our method outperforms the comparison group under all settings, illustrating that it is our designed method that unleashes the generalization ability of CLIP for FER.

Method	RAF-DB	FERPlus	AffectNet	SFEW2.0	MMA	Mean
CLIP+Finetune	88.69	59.71	44.31	41.30	44.62	55.73
CAFE	**88.72**	**73.16**	**45.86**	**52.86**	**56.80**	**63.48**
	+0.03	+13.45	+1.55	+11.56	+12.18	+7.75

5.7 Comparison with CLIP+Finetune

To demonstrate the significant generalization ability inherent in our proposed method, we construct a comparison group comprising a fixed CLIP model alongside a trainable FER model, which has similar trainable parameters with our method for fair comparison. We then fine-tune a fully connected layer on the concatenated features. The results of this comparison are depicted in Tab. 4, confirming that it is the specific design of our method that effectively unleashes CLIP's generalization potential for FER.

5.8 The Effect of the Sigmoid Function

To illustrate the effectiveness of the sigmoid function applied to the learned mask, we design a comparison group of our method without the sigmoid function, while we keep the others exactly the same. The experiment results on different FER test sets are shown in Tab. 5. The results demonstrate that the sigmoid function for mask learning is very important, without it, our method barely works. The reason lies in that the sigmoid function normalizes the learned masks, ensuring

Table 5: Influence of the sigmoid function on our method. Our method uses the sigmoid mask to select expression-related features, the comparison group has no sigmoid function and the others are the same as our method. The results show that the sigmoid function is very important for our method to learn generalizable masks.

Method	RAF-DB	FERPlus	AffectNet	SFEW2.0	MMA	Mean
No sigmoid	**89.24**	62.58	**46.37**	46.58	45.49	58.05
CAFE	88.72	**73.16**	45.86	**52.86**	**56.80**	**63.48**
	-0.52	+10.58	-0.51	+6.28	+11.31	+5.43
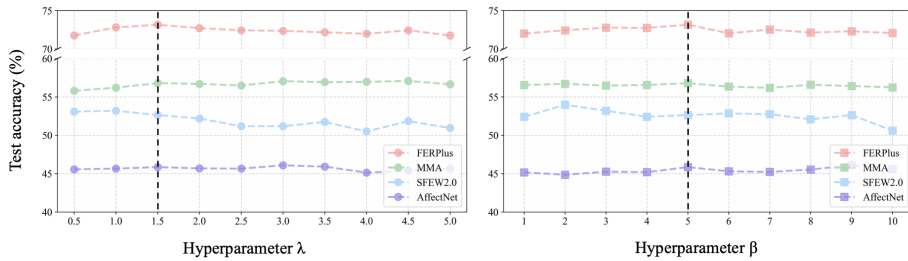
Fig. 3: The hyperparameter study of our method. Our method is not very sensitive to the two hyperparameters and we could choose them from a wide range. For simplicity, we use $\lambda = 1.5$ and $\beta = 5$ across all experiments. The cases when λ is 0 or when β is 0 are studied in Tab. 2.

that the values fall within $[0, 1]$, which reduces the overfitting ability of the learned masks.

5.9   Hyperparameter Study

We study the influence of the weight of separation loss λ and the weight of diverse loss β on our method. The results shown in Fig. 3 illustrate that both the λ and β can be chosen from a wide range and the performance is only slightly different across an order of magnitude, e.g., λ from $[0.5, 5]$ and β from $[1, 10]$. For simplicity, we choose λ as 1.5 and β as 5 across all our experiments. The cases when λ is 0 or when β is 0 are studied in Tab. 2.

5.10   Visualization Results

We use GradCAM [35] to visualize the features of FERPlus test samples by the models trained on RAF-DB. Results are shown in Fig. 4. As both EAC and our method utilize ResNet-18 instead of Transformers as the backbone, the learned features do not lie in some local areas. According to our visualization results and analyses in paper [49], EAC focuses on wrong partial features, while our method focuses on the whole face, learns better expression-related features

Fig. 4: The extracted features of FERPlus test samples by the RAF-DB trained model of EAC and our method. Labels are displayed in black, correct predictions are displayed in blue, and incorrect predictions in red.
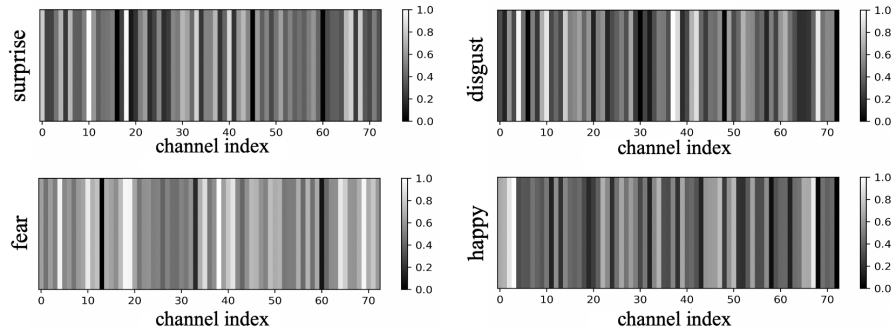

Fig. 5: Visualization of the sigmoid masks.

on unseen test samples. We also plot the learned sigmoid masks in Fig. 5, the learned masks corresponding to different expression classes show clearly different patterns, which implicitly illustrate the effectiveness of our method as some feature channels from the fixed face model are constantly useful for specific expression classes.

6   Conclusion

In this paper, we aim to improve the zero-shot generalization ability of FER methods on different unseen test sets with only one train set. Enlightened by how humans first detect faces and then recognize expressions, we propose a novel method to learn sigmoid masks on fixed general face features to extract expression-related features, combining the generalization ability of large models extracted features and the high precision of FER models. To further improve the generalization ability on unseen test sets, we propose channel-separation and channel-diverse modules to regularize the learned sigmoid masks. Extensive experiments on five different FER datasets illustrate that our method achieves the best generalization ability across SOTA FER methods.

Acknowledgements

References

1. Bargal, S.A., Barsoum, E., Ferrer, C.C., Zhang, C.: Emotion recognition in the wild from videos using images. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction. pp. 433–436 (2016)
2. Barsoum, E., Zhang, C., Ferrer, C.C., Zhang, Z.: Training deep networks for facial expression recognition with crowd-sourced label distribution. In: Proceedings of the 18th ACM international conference on multimodal interaction. pp. 279–283 (2016)
3. Carlucci, F.M., D'Innocente, A., Bucci, S., Caputo, B., Tommasi, T.: Domain generalization by solving jigsaw puzzles. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2229–2238 (2019)
4. Dhall, A., Goecke, R., Lucey, S., Gedeon, T.: Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In: 2011 IEEE international conference on computer vision workshops (ICCV workshops). pp. 2106–2112. IEEE (2011)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
6. Dou, Q., Coelho de Castro, D., Kamnitsas, K., Glocker, B.: Domain generalization via model-agnostic learning of semantic features. Advances in neural information processing systems **32** (2019)
7. Farzaneh, A.H., Qi, X.: Facial expression recognition in the wild via deep attentive center loss. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2402–2411 (2021)
8. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: International conference on machine learning. pp. 1180–1189. PMLR (2015)
9. Goodfellow, I.J., Erhan, D., Carrier, P.L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.H., et al.: Challenges in representation learning: A report on three machine learning contests. In: International conference on neural information processing. pp. 117–124. Springer (2013)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
11. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
12. Hur, S., Shin, I., Park, K., Woo, S., Kweon, I.S.: Learning classifiers of prototypes and reciprocal points for universal domain adaptation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 531–540 (2023)
13. Jia, Y., Zhang, J., Shan, S., Chen, X.: Single-side domain generalization for face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8484–8493 (2020)

14. Jin, X., Lai, Z., Jin, Z.: Learning dynamic relationships for facial expression recognition based on graph convolutional network. IEEE Transactions on Image Processing **30**, 7143–7155 (2021)
15. Kahou, S.E., Pal, C., Bouthillier, X., Froumenty, P., Gülçehre, Ç., Memisevic, R., Vincent, P., Courville, A., Bengio, Y., Ferrari, R.C., et al.: Combining modality specific deep neural networks for emotion recognition in video. In: Proceedings of the 15th ACM on International conference on multimodal interaction. pp. 543–550 (2013)
16. Kang, G., Jiang, L., Yang, Y., Hauptmann, A.G.: Contrastive adaptation network for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4893–4902 (2019)
17. Kim, D., Song, B.C.: Emotion-aware multi-view contrastive learning for facial emotion recognition. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII. pp. 178–195. Springer (2022)
18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
19. Lee, I., Lee, E., Yoo, S.B.: Latent-ofer: Detect, mask, and reconstruct with latent vectors for occluded facial expression recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1536–1546 (2023)
20. Lee, S., Kim, D., Kim, N., Jeong, S.G.: Drop to adapt: Learning discriminative features for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 91–100 (2019)
21. Li, H., Wang, N., Ding, X., Yang, X., Gao, X.: Adaptively learning facial expression representation via cf labels and distillation. IEEE Transactions on Image Processing **30**, 2016–2028 (2021)
22. Li, H., Wang, N., Yang, X., Gao, X.: Crs-cont: a well-trained general encoder for facial expression analysis. IEEE Transactions on Image Processing **31**, 4637–4650 (2022)
23. Li, H., Wang, N., Yang, X., Wang, X., Gao, X.: Unconstrained facial expression recognition with no-reference de-elements learning. IEEE Transactions on Affective Computing (2023)
24. Li, H., Wang, N., Yu, Y., Yang, X., Gao, X.: Lban-il: A novel method of high discriminative representation for facial expression recognition. Neurocomputing **432**, 159–169 (2021)
25. Li, S., Deng, W., Du, J.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2852–2861 (2017)
26. Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In: International conference on machine learning. pp. 97–105. PMLR (2015)
27. Long, M., Cao, Z., Wang, J., Jordan, M.I.: Conditional adversarial domain adaptation. Advances in neural information processing systems **31** (2018)
28. Long, M., Zhu, H., Wang, J., Jordan, M.I.: Deep transfer learning with joint adaptation networks. In: International conference on machine learning. pp. 2208–2217. PMLR (2017)
29. Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing (2017)

30. Qiao, F., Zhao, L., Peng, X.: Learning to learn single domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12556–12565 (2020)
31. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
32. Rahman, M.M., Fookes, C., Baktashmotlagh, M., Sridharan, S.: Correlation-aware adversarial domain adaptation and generalization. Pattern Recognition **100**, 107124 (2020)
33. Rangwani, H., Aithal, S.K., Mishra, M., Jain, A., Radhakrishnan, V.B.: A closer look at smoothness in domain adversarial training. In: International Conference on Machine Learning. pp. 18378–18399. PMLR (2022)
34. Ruan, D., Yan, Y., Lai, S., Chai, Z., Shen, C., Wang, H.: Feature decomposition and reconstruction learning for effective facial expression recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7660–7669 (2021)
35. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
36. Shan, C., Gong, S., McOwan, P.W.: Facial expression recognition based on local binary patterns: A comprehensive study. Image and vision Computing (2009)
37. Shankar, S., Piratla, V., Chakrabarti, S., Chaudhuri, S., Jyothi, P., Sarawagi, S.: Generalizing across domains via cross-gradient training. arXiv preprint arXiv:1804.10745 (2018)
38. Sun, B., Saenko, K.: Deep coral: Correlation alignment for deep domain adaptation. In: Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14. pp. 443–450. Springer (2016)
39. Sun, T., Lu, C., Zhang, T., Ling, H.: Safe self-refinement for transformer-based domain adaptation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7191–7200 (2022)
40. Tang, K., Niu, Y., Huang, J., Shi, J., Zhang, H.: Unbiased scene graph generation from biased training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3716–3725 (2020)
41. Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., Darrell, T.: Deep domain confusion: Maximizing for domain invariance. arXiv preprint arXiv:1412.3474 (2014)
42. Wang, K., Peng, X., Yang, J., Lu, S., Qiao, Y.: Suppressing uncertainties for large-scale facial expression recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6897–6906 (2020)
43. Wang, K., Peng, X., Yang, J., Meng, D., Qiao, Y.: Region attention networks for pose and occlusion robust facial expression recognition. IEEE Transactions on Image Processing (2020)
44. Xu, T., Chen, W., Wang, P., Wang, F., Li, H., Jin, R.: Cdtrans: Cross-domain transformer for unsupervised domain adaptation. arXiv preprint arXiv:2109.06165 (2021)
45. Yang, J., Liu, J., Xu, N., Huang, J.: Tvt: Transferable vision transformer for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 520–530 (2023)

46. Yang, S., Jui, S., van de Weijer, J., et al.: Attracting and dispersing: A simple approach for source-free domain adaptation. Advances in Neural Information Processing Systems **35**, 5802–5815 (2022)
47. Zeng, D., Lin, Z., Yan, X., Liu, Y., Wang, F., Tang, B.: Face2exp: Combating data biases for facial expression recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20291–20300 (2022)
48. Zhang, Y., Wang, C., Deng, W.: Relative uncertainty learning for facial expression recognition. Advances in Neural Information Processing Systems **34**, 17616–17627 (2021)
49. Zhang, Y., Wang, C., Ling, X., Deng, W.: Learn from all: Erasing attention consistency for noisy label facial expression recognition. In: European Conference on Computer Vision. pp. 418–434. Springer (2022)
50. Zhi, R., Flierl, M., Ruan, Q., Kleijn, W.B.: Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) **41**(1), 38–52 (2010)
51. Zhong, L., Liu, Q., Yang, P., Liu, B., Huang, J., Metaxas, D.N.: Learning active facial patches for expression analysis. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 2562–2569. IEEE (2012)
52. Zhou, K., Yang, Y., Hospedales, T., Xiang, T.: Deep domain-adversarial image generation for domain generalisation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 13025–13032 (2020)
53. Zhou, K., Yang, Y., Hospedales, T., Xiang, T.: Learning to generate novel domains for domain generalization. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16. pp. 561–578. Springer (2020)