When Facial Expression Recognition Meets Few-Shot Learning: A Joint and Alternate Learning Framework

Xinyi Zou[1], **Yan Yan**[1*], **Jing-Hao Xue**[2], **Si Chen**[3], **Hanzi Wang**[1]

[1]Xiamen University, China [2]University College London, UK [3]Xiamen University of Technology, China

Abstract

Human emotions involve basic and compound facial expressions. However, current research on facial expression recognition (FER) mainly focuses on basic expressions, and thus fails to address the diversity of human emotions in practical scenarios. Meanwhile, existing work on compound FER relies heavily on abundant labeled compound expression training data, which are often laboriously collected under the professional instruction of psychology. In this paper, we study compound FER in the cross-domain few-shot learning setting, where only a few images of novel classes from the target domain are required as a reference. In particular, we aim to identify unseen compound expressions with the model trained on easily accessible basic expression datasets. To alleviate the problem of limited base classes in our FER task, we propose a novel Emotion Guided Similarity Network (EGS-Net), consisting of an emotion branch and a similarity branch, based on a two-stage learning framework. Specifically, in the first stage, the similarity branch is jointly trained with the emotion branch in a multi-task fashion. With the regularization of the emotion branch, we prevent the similarity branch from overfitting to sampled base classes that are highly overlapped across different episodes. In the second stage, the emotion branch and the similarity branch play a "two-student game" to alternately learn from each other, thereby further improving the inference ability of the similarity branch on unseen compound expressions. Experimental results on both in-the-lab and in-the-wild compound expression datasets demonstrate the superiority of our proposed method against several state-of-the-art methods.

Introduction

Over the past few decades, facial expression recognition (FER) has attracted considerable attention because of its wide range of applications in human-robot interaction, online education, driver monitoring, *etc* (Corneanu et al. 2016).

Based on Ekman and Friesen's study (Ekman and Friesen 1971), facial expressions are typically classified into seven basic expressions, including happiness, sadness, disgust, anger, fear, surprise, and neutral. Previous work on FER chiefly focuses on the classification of these pre-defined basic expressions. Accordingly, numerous basic expression datasets (Lucey et al. 2010; Li, Deng, and Du 2017; Zhao et al. 2011) have been collected, and impressive progress (Li et al. 2018; Ruan et al. 2020; Zhao, Liu, and Zhou 2021) has been made to address large facial appearance variations caused by identity, pose, occlusion, illumination, and so on. In this paper, we refer to the above conventional FER task as the basic FER task.

Regrettably, these basic expressions cannot completely characterize the diversity of human emotions in nature. Du *et al.* (Du, Tao, and Martinez 2014) reveal that human emotions involve compound expressions, beyond the above basic expressions. They enlarge the number of expressions to 22 by combining basic expressions. Later, the EmotioNet dataset (Fabian Benitez-Quiroz, Srinivasan, and Martinez 2016) is constructed with large-scale compound expression data. To classify the above compound expressions, conventional deep learning based methods (Slimani et al. 2019; Guo et al. 2017) usually rely heavily on a large amount of labeled compound expression training data. However, collecting such data is laborious and often demands the professional instruction of psychology.

As humans, given only a few reference images (a support set), we can easily recognize an unseen expression (a query) based on the prior knowledge of various seen expressions. Recent research on few-shot learning (FSL) exhibits the potential of quickly generalizing to novel classes with only a few labeled data of these classes, thereby reducing the gap between humans and artificial intelligence (Lu et al. 2020). In this paper, we investigate compound FER in the cross-domain FSL (CD-FSL) paradigm, which greatly alleviates the burden of collecting large-scale labeled compound expression data. Notably, instead of manually splitting a compound expression dataset into a base class set and a novel class set, we explore a more challenging but practical setup, which aims to classify compound expressions from the unseen domain by using the model trained only on easily accessible basic expression datasets.

Traditional FSL methods have achieved promising performance in many computer vision tasks, such as image classification (Li et al. 2019; Yao et al. 2020) and object detection (Dong et al. 2018; Yang et al. 2020). However, few work is concerned with the compound FER task in the CD-FSL setting. Different from widely used FSL benchmarks (e.g., miniImageNet (Vinyals et al. 2016) and Ominiglot (Lake,

*Corresponding author (email: yanyan@xmu.edu.cn).

Salakhutdinov, and Tenenbaum 2015) whose total numbers of classes are 100 and 1,623, respectively), basic expression datasets contain a limited number of basic expressions (i.e., base classes in our setting). Consequently, the random sampling process cannot effectively simulate the variance of unseen tasks since the sampled base classes are highly overlapped across different episodes. In this way, traditional FSL methods easily suffer from the overfitting problem, resulting in their deteriorated inference ability on unseen compound expressions.

To address the above problem, we propose an effective CD-FSL method called Emotion Guided Similarity Network (EGS-Net), consisting of an emotion branch and a similarity branch, for compound FER. The emotion branch captures the global information of basic expressions and serves as a regularizer, while the similarity branch learns a transferable similarity metric between two expressions. In particular, motivated by the human perception that one can better identify compound expressions with more prior knowledge of basic expressions, we develop a two-stage learning framework to train EGS-Net in a progressive manner: (i) joint learning of the emotion branch and the similarity branch in a multi-task fashion; (ii) alternate learning between the emotion branch and the similarity branch. As a result, our proposed method remarkably relieves the requirement of abundant compound expression training data and offers superior scalability for practical applications.

In summary, our main contributions are given as follows:

- We propose a novel EGS-Net method for compound FER in the CD-FSL setting. Our method is capable of learning a transferable model, which is trained only on multiple basic expression datasets. Therefore, we can easily recognize novel compound expressions from the unseen domain, with a few reference images of novel classes. To the best of our knowledge, we are the first to classify unseen compound expressions in the FSL scenario.
- We develop a two-stage learning framework to progressively train EGS-Net and thus effectively alleviate the problem of limited base classes in our FER task. Based on the proposed learning framework, the inference ability of the similarity branch can be greatly improved with the help of the emotion branch, thereby boosting the performance of predicting novel compound expressions.
- Extensive experimental results on both in-the-lab and in-the-wild compound expression datasets demonstrate the effectiveness of our proposed method in comparison with several state-of-the-art FSL methods.

Related Work

Facial Expression Recognition. The past decades have witnessed significant progress in FER. Considering its practical applications, the main focus of FER has shifted from controllable in-the-lab scenarios to more challenging in-the-wild scenarios. However, conventional FER methods (Li et al. 2018; Ruan et al. 2020; Zhao, Liu, and Zhou 2021) only classify basic expressions, and fail to depict the complexity of human emotions in practical scenarios.

Recently, Du *et al.* (Du, Tao, and Martinez 2014) reveal that there are a large number of emotions expressed regularly by humans. They further define the compound expressions by combining basic expressions. Benitez-Quiroz *et al.* (Fabian Benitez-Quiroz, Srinivasan, and Martinez 2016) introduce a large compound expression dataset called EmotioNet, which contains one million in-the-wild images labeled by an AU-based algorithm. Based on the above datasets, several attempts are made for compound FER. Slimani *et al.* (Slimani et al. 2019) propose a highway convolutional neural network which replaces the shortcut with a learnable parameter for compound FER. As a winner of the FG 2017 Challenge, Guo *et al.* (Guo et al. 2017) design a multi-modality convolutional neural network, which combines the visual feature with the geometry feature and shows superiority for the emotion challenge.

Conventional compound FER methods require a large amount of labeled compound expression training data. Collecting such data not only is time-consuming and labor-intensive, but also demands the professional guidance. In this paper, different from the above methods, we are concerned with the compound FER problem in the CD-FSL setting, where the base classes are sampled from multiple basic expression datasets and the novel classes are compound expressions. Therefore, we manage to perform compound FER with only a few labeled reference images and provide great flexibility to identify a new expression category.

Few-Shot Learning. With the success of convolutional neural networks, deep learning based FSL methods have become topical. These methods can be coarsely classified into meta-learning based methods (Vinyals et al. 2016; Snell, Swersky, and Zemel 2017; Sung et al. 2018; Garcia and Bruna 2018; Finn, Abbeel, and Levine 2017) and transfer learning based methods (Chen et al. 2019; Afrasiyabi, Lalonde, and Gagné 2020; Hu, Gripon, and Pateux 2021; Yang, Liu, and Xu 2021). In this paper, our method belongs to meta-learning based methods and it is based on the learn-to-measure (L2M) technique that aims to learn a transferable similarity metric.

Recently, some FSL methods (Luo et al. 2017; Tseng et al. 2020) are also developed under the cross-domain setting. For example, Luo *et al.* (Luo et al. 2017) adopt adversarial learning to learn a transferable representation across different domains. Tseng *et al.* (Tseng et al. 2020) propose novel feature-wise transformation layers to simulate the variance of the target domain. Guo *et al.* (Guo et al. 2020) investigate a more challenging scenario, where a large domain shift exists between the base class domain and the novel class one.

Although existing FSL methods have shown promising performance in a variety of computer vision tasks, few of them study the compound FER task. The most relevant work to ours is (Ciubotaru et al. 2019), which evaluates some representative FSL methods for the basic FER task rather than generalizing to classify unseen compound expressions. In fact, due to the limited number of base classes in our FER task, the performance of existing FSL methods drops substantially. Hence, we develop a novel Emotion Guided Similarity Network (EGS-Net) based on a two-stage learning framework to address this issue.

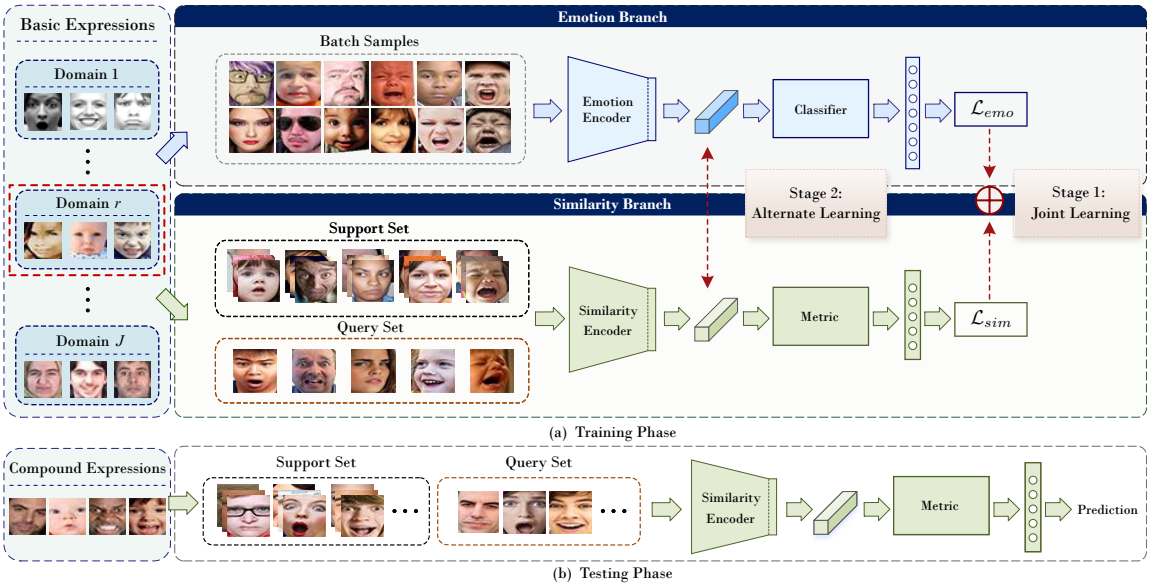

Figure 1: Overview of the proposed EGS-Net, which consists of an emotion branch and a similarity branch. (a) During the training phase, EGS-Net is progressively trained by using a two-stage learning framework. In stage 1, we perform joint learning of the two branches in a multi-task fashion. In stage 2, we perform alternate learning between the two branches. (b) During the testing phase, the performance is evaluated on the compound expression dataset based on the learned similarity branch.

Proposed Method

Problem Definition

In this paper, we perform compound FER in the CD-FSL setting, where the classes of the training set (i.e., the base class set) and those of the test set (i.e., the novel class set) are disjoint, and they are from different domains. To enrich the diversity of base classes and bridge the domain gap between the training set and the test set, we adopt multiple source domains (i.e., multiple basic expression datasets) for training. Accordingly, basic expressions from source domains are introduced to construct the base classes and compound expressions from the target domain (i.e., a compound expression dataset) are used as the novel classes. Such a setting is a challenging but practical setup, which investigates the ability of recognizing novel compound expressions based on the model trained only on easily accessible basic expression datasets. Therefore, given a base class set with sufficient labeled images, we aim to learn a transferable model and evaluate its performance on a novel class set with a few reference images. This enables the flexibility of the model to address compound FER.

Overview

An overview of the proposed Emotion Guided Similarity Network (EGS-Net) is shown in Figure 1. EGS-Net consists of an emotion branch and a similarity branch. The emotion branch learns global feature representations to classify all the basic expressions, while the similarity branch learns a transferable similarity metric between two expressions. Specifically, for the training phase, the emotion branch is learned by mini-batch training. Meanwhile, the similarity branch follows the L2M setting and it is trained in an episodic manner. In each episode, a meta-task is performed by sampling a support set and a query set from a randomly selected source domain, and then the model parameters are updated by the classification errors on the sampled query set. For the testing phase, we construct similar meta-tasks from the compound expression dataset. In each meta-task, based on the learned similarity branch, a query image is classified into its nearest category in the support set.

In particular, considering the difficulty of performing compound FER in the CD-FSL setting due to the limited base classes, a two-stage learning framework (including a joint learning stage and an alternate learning stage) is developed to train EGS-Net progressively. In the first stage, the emotion branch and the similarity branch are jointly trained in a multi-task fashion. In the second stage, the two branches are separately trained by alternate learning. They are updated alternately with the guidance of each other. As a result, the learned similarity branch can better serve for the unseen compound FER task, given only a few reference images of novel classes.

Joint Learning

Different from conventional FSL benchmarks that contain a large number of classes, basic expression datasets involve only a few basic expression categories (i.e., base classes in our setting). As a consequence, the constructed few-shot classification tasks are severely overlapped across different episodes, and existing FSL methods are likely to be trapped into the sampled base classes, leading to overfitting. To address this problem, we jointly train the emotion branch and the similarity branch. During the joint learning stage, the emotion branch, which captures the global information of

basic expressions, is served as a regularizer to avoid overfitting of the similarity branch. Such a way significantly improves the inference ability of the similarity branch on basic expressions from the unseen domain, and thus facilitates the training of the second stage. The optimization objective of this stage is formulated as

$$\mathcal{L}_{joint} = \mathcal{L}_{sim} + \lambda_{emo}\mathcal{L}_{emo}, \tag{1}$$

where $\mathcal{L}_{joint}$ denotes the joint loss. $\mathcal{L}_{sim}$ and $\mathcal{L}_{emo}$ represent the classification losses of the similarity branch and the emotion branch, respectively. λ_{emo} denotes the balanced parameter.

In the following, we will introduce the emotion branch and the similarity branch in detail.

Emotion Branch. The emotion branch consists of an emotion encoder E_e and a classifier f to classify the basic expressions. By performing the basic FER task, the emotion branch provides a global view of all basic expression information. Given multiple source domains $\mathbb{D}_s = \{D_1, D_2, \cdots, D_J\}$, where D_j represents the j-th source domain and J is the total number of training domains, a source domain D_r is randomly selected in every episode. The batch data $\{X_i^r, Y_i^r\}$ are sampled from D_r, where X_i^r and Y_i^r denote the batch images and their corresponding labels, respectively. The predicted label $\hat{y}_i^r$ is then computed as $\hat{y}_i^r = f(E_e(\boldsymbol{x}_i^r))$, where $\boldsymbol{x}_i^r$ denotes a single image from the sampled batch. The classification loss of the emotion branch $\mathcal{L}_{emo}$ employs the popular cross-entropy loss between the predicted result $\hat{y}_i^r$ and the ground-truth expression label y_i^r, that is,

$$\mathcal{L}_{emo} = -\sum_{c=1}^{C_r} \mathbb{1}_{[c=y_i^r]} \log(f(E_e(\boldsymbol{x}_i^r))), \tag{2}$$

where C_r denotes the number of basic expression categories in D_r. Indicator function $\mathbb{1}_{[c=y_i^r]}$ equals to 1 only if $c = y_i^r$, and 0 otherwise.

Similarity Branch. The similarity branch involves a similarity encoder E_s and a metric module M. The similarity encoder E_s and the emotion encoder E_e share the parameters in the joint learning stage. Mathematically, for a meta-training episode, given a randomly selected domain D_r, the training data are randomly sampled and divided into a support set $\mathbb{S} = \{X_s^r, Y_s^r\}$ and a query set $\mathbb{Q} = \{X_q^r, Y_q^r\}$, where X_s^r, Y_s^r and X_q^r, Y_q^r denote the sampled images and their corresponding labels in the support set and the query set, respectively. Subsequently, an N-way K-shot classification task is constructed, where N denotes the number of sampled classes and K represents the number of the labeled images in each class of the support set. The goal of a few-shot classification task is to make predictions for the query images with the reference of the support set.

All the images from the support set and the query set are fed into the similarity branch to evaluate the similarity between them. Then, the query image is assigned to its nearest category according to the similarity between this image and the support set in the learned feature space. The prediction process is formulated as

$$\hat{Y}_q^r = g(M(E_s(X_s^r), E_s(X_q^r)), Y_s^r), \tag{3}$$

where $\hat{Y}_q^r$ represents the predicted results for the query images. $M(\cdot)$ denotes the metric function, and $g(\cdot)$ refers to the operation that assigns a query image to its nearest category according to the similarity metric.

The objective of each few-shot classification task is to minimize the loss between the predicted result $\hat{y}_q^r$ and the ground-truth label y_q^r of each query image as

$$\mathcal{L}_{sim} = -\sum_{n=1}^{N} \mathbb{1}_{[n=y_q^r]} \log(\hat{y}_q^r). \tag{4}$$

By training across different meta-tasks, the similarity branch can be easily adapted to an unseen task.

Alternate Learning

After joint learning, the inference ability of the similarity branch on basic expressions from the unseen domain is substantially improved, while that on compound expressions is still inferior. This is due to the poor inference ability of the initial emotion branch on novel classes. Motivated by the observation that humans are able to better learn knowledge by communicating with each other from a different perspective, we further develop the alternate learning stage. This learning stage can be viewed as a "two-student game", where one student (branch) learns from the other one in turn.

More specifically, at the beginning of this stage, we update the emotion branch to perform its own expression classification task under the supervision of the fixed similarity branch for K_e periods. Given a sampled image $\boldsymbol{x}_i^r$, the objective function $\mathcal{L}_{emo}^{al}$ in this step is given as

$$\mathcal{L}_{emo}^{al} = \mathcal{L}_{emo} + \theta_{n_e} ||E_s(\boldsymbol{x}_i^r) - E_e(\boldsymbol{x}_i^r)||_2^2, \tag{5}$$

where $\mathcal{L}_{emo}$ denotes the classification loss defined in Eq. (2) for the emotion branch. θ_{n_e} denotes the dynamic weight that varies with the episode n_e. In this paper, we adopt a weight decay strategy to highlight the key role of the emotion branch in this step during alternate learning. $||E_s(\boldsymbol{x}_i^r) - E_e(\boldsymbol{x}_i^r)||_2$ is a regularized term, which constrains the feature distance between the similarity encoder and the emotion encoder to be as close as possible. Consequently, the emotion branch captures the knowledge that can be transferred to an unseen task to some extent.

Then, the role of each branch is exchanged, where the similarity branch intends to learn from the updated emotion branch in an episodic manner for K_s periods. By resorting to the enhanced inference ability, the updated emotion branch can boost the classification performance of the similarity branch on both basic and compound expressions from the unseen domain. The objective function $\mathcal{L}_{sim}^{al}$ in this step is formulated as

$$\mathcal{L}_{sim}^{al} = \mathcal{L}_{sim} + \theta_{n_e} ||E_e(\boldsymbol{x}_i^r) - E_s(\boldsymbol{x}_i^r)||_2^2, \tag{6}$$

where $\mathcal{L}_{sim}$ represents the metric based classification loss defined in Eq. (4) for the similarity branch. Similarly, we emphasize the key role of the similarity branch by the dynamic weight θ_{n_e} in this step.

Next, the similarity branch and the emotion branch are alternately trained several times. Unlike the "two-player

game" of GAN (Goodfellow et al. 2014), in which the generator and the discriminator compete with each other, the proposed alternate learning stage improves the inference ability of both branches by exchanging their respective knowledge. Finally, a similarity branch, which has superior inference ability on novel classes, can be obtained and transferred to perform the unseen compound FER task.

Overall Training

In the first stage, we jointly train the similarity branch and the emotion branch, preventing the similarity branch from overfitting to highly overlapped sampled base classes. In the second stage, we alternately train one branch with the guidance of the other one. The similarity branch is first fixed while the emotion branch is updated to improve its inference ability. Then, the similarity branch is optimized under the supervision of the updated emotion branch to better exploit the global information of basic expressions. Finally, the above two steps are alternately trained. In this way, the two branches can learn from each other from a different perspective, greatly improving the inference ability of the similarity branch to identify unseen compound expressions. The two-stage learning framework is summarized in the appendix.

Experiments

Datasets

In this paper, we study the compound FER task in the CD-FSL setting, where only images from easily accessible basic expression datasets are used to train the model. We use several popular basic expression datasets, including three in-the-lab datasets (CK+ (Lucey et al. 2010), MMI (Pantic et al. 2005), and Oulu-CASIA (Zhao et al. 2011)) and two in-the-wild datasets (RAF-DB (Li, Deng, and Du 2017) and SFEW (Dhall et al. 2011)), as multiple source domains for training. We use two newly released compound expression datasets (CFEE (Du, Tao, and Martinez 2014) and EmotioNet (Benitez-Quiroz et al. 2017)) for testing. More information of these datasets is provided in the appendix.

To better analyze the inference ability of our method, we divide CFEE into two subsets, including 1,610 images labeled with basic expressions (denoted CFEE_B) and 3,450 images labeled with compound expressions (denoted CFEE_C). Similar to CFEE, we divide EmotioNet into EmotioNet_B (consisting of basic expressions) and EmotioNet_C (consisting of compound expressions).

Implementation Details

For all the experiments, we first align and crop facial images by MTCNN (Zhang et al. 2016), and further resize them to 224×224. All the images in basic expression datasets are used for training. The compound expression datasets and their corresponding subsets are used for testing.

We implement our model with the Pytorch toolbox. We adopt ResNet-18 (He et al. 2016) as the backbone for both the emotion encoder and the similarity encoder, which share the parameters in the joint learning stage and are separately updated in the alternate learning stage. The networks are optimized by using the Adam algorithm (Kingma and Ba 2015) with the learning rate of 0.001, $\beta_1 = 0.500$, and $\beta_2 = 0.999$. The weight of the emotion branch is empirically set to $\lambda_{emo} = 1$ during the joint learning stage. We adopt the step decay strategy during the alternate learning stage. For the emotion branch, the batch size is set to 128. For the similarity branch, we randomly sample N ($= 5$) classes and K ($= 1, 5$) images from each class to form the support set, and the number of query images is set to 16. The whole training contains 200 epochs for joint learning and 5 epochs for alternate learning, and the two branches exchange the role after every 20 periods ($K_e = K_s = 20$). The number of episodes N_e in each epoch is set to 100. We report the average recognition accuracy on 1000 meta-test tasks.

Ablation Studies

To better analyze the inference ability on the unseen domain, we test the method on the whole dataset and two subsets, including a subset of basic expressions (CFEE_B or EmotioNet_B) and a subset of unseen compound expressions (CFEE_C or EmotioNet_C). The whole dataset is used to evaluate the overall accuracy, while the subsets of basic and compound expressions are used to evaluate the inference ability of a method on the seen classes and the novel classes from the unseen domain, respectively. The classical ProtoNet (Snell, Swersky, and Zemel 2017) is used as our similarity branch in this subsection.

Influence of Emotion Branch and Similarity Branch. We first evaluate the inference ability of the emotion branch and the similarity branch, when they are independently trained without using the two-stage learning framework. The two branches are evaluated in an FSL manner. That is, images from the support and query sets are fed into the trained feature extractor to extract features. The query images are then assigned to their nearest neighbors in the support set. We respectively denote the emotion branch and the similarity branch trained on a single domain (RAF-DB is used) as E_b (single) and S_b (single), and those trained on multiple source domains as E_b (multiple) and S_b (multiple). The comparison results are given in Table 1.

As illustrated in Table 1, S_b (single) and E_b (single) achieve similar performance for basic expression recognition on the unseen domains (i.e., CFEE_B and EmotioNet_B). In contrast, S_b (single) outperforms E_b (single) by a large margin for classifying unseen compound expressions (i.e., 7.69%, 6.94% on CFEE_C, and 2.90%, 2.88% on EmotioNet_C for 5-shot and 1-shot classification tasks, respectively). Similar patterns can be observed when multiple source domains are used. These results indicate that the inference ability of the similarity branch on the unseen task is better than that of the emotion branch. This can be ascribed to the superiority of the episodic training manner for the similarity branch.

Moreover, E_b (multiple) and S_b (multiple) obtain much better recognition accuracy than E_b (single) and S_b (single), respectively, on the whole datasets and their corresponding subsets. Therefore, multiple source domains effectively enrich the diversity of the training data, and bridge the gap between the training set and the test set. In the following part, we will use multiple source domains as the training set.

Method	CFEE		CFEE_B		CFEE_C		EmotioNet		EmotioNet_B		EmotioNet_C	
	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot
E_b (single)	59.36	47.48	71.57	59.87	55.00	43.13	54.86	44.01	63.18	50.24	54.76	45.25
E_b (multiple)	65.59	52.65	80.28	67.48	60.66	47.48	56.03	45.17	64.45	51.42	55.90	46.23
S_b (single)	65.41	54.22	71.44	65.73	62.69	50.07	56.35	46.38	63.13	51.12	57.66	48.13
S_b (multiple)	69.69	58.05	82.21	72.51	66.84	54.30	57.49	48.58	68.24	56.66	58.40	49.93
EGS-Net (joint)	70.88	59.18	85.63	76.74	67.05	54.99	58.57	49.14	70.60	59.16	58.83	50.57
EGS-Net (al)	71.25	60.02	84.09	75.11	67.33	55.13	58.73	49.28	69.39	57.32	59.25	50.86
EGS-Net	**72.17**	**60.90**	**86.45**	**77.16**	**68.38**	**56.65**	**59.77**	**50.06**	**71.65**	**59.67**	**60.52**	**51.62**

Table 1: The 5-shot and 1-shot accuracy (%) on CFEE, EmotioNet, and the corresponding subsets.

Method	CFEE	CFEE_C	EmotioNet	EmotioNet_C
E_b (joint)	69.14	65.46	57.61	57.44
E_b (two-stage)	**71.30**	**66.77**	**58.72**	**59.25**

Table 2: Inference ability of the emotion branch. The 5-shot accuracy (%) is reported for performance evaluation.

Weight decay	CFEE	CFEE_C	EmotioNet	EmotioNet_C
×	71.31	67.03	59.33	59.73
✓	**72.17**	**68.38**	**59.77**	**60.52**

Table 3: Influence of the weight decay strategy. The 5-shot accuracy (%) is reported for performance evaluation.

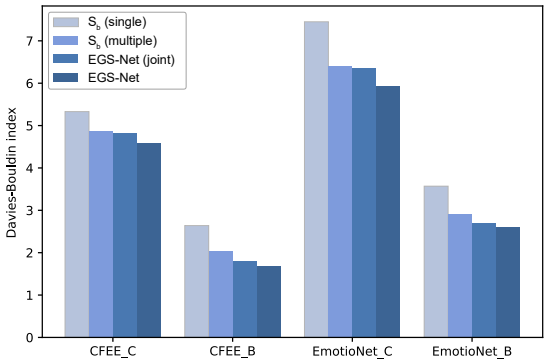

Figure 2: DB index of the learned features for basic and compound expressions on both in-the-lab and in-the-wild datasets. For the DB index, the smaller is better.

Influence of Joint Learning. The results obtained by EGS-Net only using the joint learning stage (denoted EGS-Net (joint)) are shown in Table 1. We also compare EGS-Net only using the alternate learning stage (denoted EGS-Net (al)) with EGS-Net using the two-stage learning framework.

Compared with S_b (multiple), EGS-Net (joint) obtains higher recognition accuracy on the basic expression subsets. To be specific, EGS-Net (joint) improves the performance by 3.42%, 4.23% on CFEE_B, and 2.36%, 2.50% on EmotioNet_B for the 5-shot and 1-shot classification tasks, respectively. Therefore, the joint learning stage is beneficial to alleviate the overfitting problem on sampled base classes, and thus enable EGS-Net to classify basic expressions from the unseen domain more accurately. In addition, compared with EGS-Net (al), EGS-Net respectively achieves the improvements of 1.05% and 1.27% in terms of recognition accuracy on CFEE_C and EmotioNet_C for the 5-shot classification task. This validates the necessity of the joint learning stage, which facilitates the training of the second stage.

We also compute the Davies-Bouldin index (DB index) (Davies and Bouldin 1979) of different methods on four subsets. DB index depicts the intra-class variations and the inter-class similarities in the learned feature space. For the DB index, the smaller is better. The results are shown in Figure 2. We can observe that the DB index obtained by EGS-Net (joint) is better than that obtained by S_b (multiple) on all subsets. However, the DB index decreases more on basic expression subsets (i.e., CFEE_B and EmotioNet_B) than that on novel compound subsets (i.e., CFEE_C and EmotioNet_C). This shows that the inference ability of EGS-Net (joint) on unseen compound expressions is still inferior. The main reason is that the poor inference ability of the initial E_b (multiple) on the unseen task constrains the performance of the similarity branch during joint learning.

Influence of Alternate Learning. As shown in Table 1, EGS-Net (al) gives higher recognition accuracy than S_b (multiple) on CFEE, EmotioNet, and their corresponding subsets. Compared with EGS-Net (joint), EGS-Net (al) gives better accuracy on the compound subsets but performs worse on the basic subsets. This is because the alternate learning stage facilitates our model to identify unseen compound expressions by training the similarity branch across similar tasks. However, EGS-Net (al) still suffers from the overfitting problem caused by limited base classes.

The two-stage EGS-Net further improves the performance of EGS-Net (joint), especially for the unseen compound FER task. Specifically, it obtains improvements of 1.33%, 1.66% on CFEE_C, and 1.69%, 1.05% on EmotioNet_C for the 5-shot and 1-shot classification tasks, respectively. Hence, the alternate learning stage is of great significance to enhance the inference ability of the similarity branch.

Moreover, we also evaluate the inference ability of the emotion branch, as shown in Table 2. We give the results obtained by the emotion branch based only on the joint learning stage (denoted E_b (joint)) and that based on the two-stage learning framework (denoted E_b (two-stage)). We can see that the inference ability of the emotion branch on the unseen task is enhanced after the alternate learning stage (1.31% and 1.81% improvements on CFEE_C and EmotioNet_C, respectively). Resorting to the improved inference

Method	CFEE		EmotioNet	
	5-shot	1-shot	5-shot	1-shot
ProtoNet (Snell, Swersky, and Zemel 2017)	69.69	58.05	57.49	48.58
MatchingNet (Vinyals et al. 2016)	64.70	56.75	54.14	48.09
RelationNet (Sung et al. 2018)	65.27	56.51	56.18	48.33
GNN (Garcia and Bruna 2018)	70.10	58.45	58.06	49.23
InfoPatch (Liu et al. 2021)	71.99	60.82	58.73	46.61
DKT (Patacchiola et al. 2020)	67.55	54.94	55.30	45.39
GNN+LFT (Tseng et al. 2020)	71.76	59.96	61.37	51.56
BASELINE (Chen et al. 2019)	66.98	54.21	60.15	50.38
BASELINE++ (Chen et al. 2019)	68.60	56.28	61.13	51.00
Arcmax loss (Afrasiyabi, Lalonde, and Gagné 2020)	68.92	56.94	60.87	51.02
PT+NCM (Hu, Gripon, and Pateux 2021)	68.59	56.60	55.70	46.45
LR+DC (Yang, Liu, and Xu 2021)	68.97	57.97	55.71	46.98
EGS-Net (P)	72.17	60.90	59.77	50.06
EGS-Net (M)	67.43	58.06	56.24	49.21
EGS-Net (R)	67.28	57.60	56.90	49.55
EGS-Net (G)	**73.79**	**61.28**	**62.12**	**51.93**

Table 4: The 5-shot and 1-shot accuracy (%) comparisons among different competing methods on the in-the-lab CFEE and in-the-wild EmotioNet datasets.

ability, the emotion branch can better guide the training of the similarity branch in the second learning stage.

In this paper, a weight decay strategy is introduced to highlight the key role of current training branch during the alternate learning stage. The influence of the weight decay strategy is shown in Table 3. We can observe that the weight decay strategy is beneficial to improve the performance.

Finally, we demonstrate the discriminability of the learned features obtained by EGS-Net. From Figure 2, EGS-Net gives better DB index than EGS-Net (joint) on four subsets. The gap is more evident on unseen compound expression subsets. This further validates the importance of the proposed alternate learning stage. Furthermore, we also show some feature visualization results in the appendix.

Comparison with State-of-the-Art Methods

Table 4 gives the comparison results between our proposed method and several state-of-the-art FSL methods on the compound expression datasets. We build our EGS-Net methods based on four representative L2M methods, including ProtoNet (Snell, Swersky, and Zemel 2017), MatchingNet (Vinyals et al. 2016), RelationNet (Sung et al. 2018), and GNN (Garcia and Bruna 2018), denoted EGS-Net (P), EGS-Net (M), EGS-Net (R), and EGS-Net (G), respectively. These methods differ in terms of metric modules. Specifically, ProtoNet and MatchingNet employ the Euclidean and cosine distances, respectively. A learnable metric module based on the vanilla and graph convolution is used in RelationNet and GNN. For a fair comparison, all the competing methods are trained by using publicly available codes under the same settings (e.g., dataset and backbone).

Compared with the corresponding L2M baselines, EGS-Net (P), EGS-Net (M), EGS-Net (R), and EGS-Net (G) achieve higher performance (2.48%, 2,73%, 2.01%, 3.69% improvements on CFEE, and 2.28%, 2.10%, 0.72%, 4.06% improvements on the more challenging EmotioNet dataset for the 5-shot classification tasks). The above results indicate that our proposed EGS-Net method can further improve the inference ability of existing L2M methods on the unseen compound expression datasets.

Moreover, we evaluate several recent FSL methods for performance comparison. For instance, InfoPatch (Liu et al. 2021) introduces contrastive learning into the episodic training manner for a general matching. DKT (Patacchiola et al. 2020) learns a kernel that can transfer to a new task for the Bayesian model. Tseng *et al.* (Tseng et al. 2020) solve the CD-FSL problem by using feature-wise transformation layers. Some transfer learning based methods focus on either the design of loss functions in the pretraining stage (Chen et al. 2019; Afrasiyabi, Lalonde, and Gagné 2020) or the calibration of novel class distribution in the fine-tuning stage (Hu, Gripon, and Pateux 2021; Yang, Liu, and Xu 2021). As can be seen in Table 4, among all the competing methods, our EGS-Net (G), which uses a graph convolution based metric function, obtains the highest accuracy of 73.79%, 61.28% on the in-the-lab CFEE dataset, and 62.12%, 51.93% on the in-the-wild EmotioNet dataset for 5-shot and 1-shot classification tasks, respectively.

Conclusion

In this paper, we propose a novel EGS-Net method for compound FER in the CD-FSL setting, which substantially avoids the tedious collection of large-scale labeled compound expression training data and offers superior scalability for practical applications. To alleviate the problem of limited base classes, a novel two-stage learning framework is developed. The proposed framework includes a joint learning stage to prevent the trained model from overfitting to highly overlapped sampled base classes, and an alternate learning stage to further improve the inference ability of our model for generalizing to the unseen task. Extensive experiments have been performed to validate the effectiveness of our method on both in-the-lab and in-the-wild compound expression datasets.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grants 62071404 and 61872307, by the Open Research Projects of Zhejiang Lab under Grant 2021KG0AB02, by the Natural Science Foundation of Fujian Province under Grant 2020J01001, and by the Youth Innovation Foundation of Xiamen City under Grant 3502Z20206046.

References

Afrasiyabi, A.; Lalonde, J.-F.; and Gagné, C. 2020. Associative alignment for few-shot image classification. In *Proceedings of the European Conference on Computer Vision*, 18–35.

Benitez-Quiroz, C. F.; Srinivasan, R.; Feng, Q.; Wang, Y.; and Martinez, A. M. 2017. EmotioNet challenge: Recognition of facial expressions of emotion in the wild. *arXiv preprint arXiv:1703.01210*.

Chen, W.-Y.; Liu, Y.-C.; Kira, Z.; Wang, Y.-C.; and Huang, J.-B. 2019. A closer look at few-shot classification. In *Proceedings of the International Conference on Learning Representations*.

Ciubotaru, A.-N.; Devos, A.; Bozorgtabar, B.; Thiran, J.-P.; and Gabrani, M. 2019. Revisiting few-shot learning for facial expression recognition. *arXiv preprint arXiv:1912.02751*.

Corneanu, C. A.; Simón, M. O.; Cohn, J. F.; and Guerrero, S. E. 2016. Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8): 1548–1568.

Davies, D. L.; and Bouldin, D. W. 1979. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2): 224–227.

Dhall, A.; Goecke, R.; Lucey, S.; and Gedeon, T. 2011. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2106–2112.

Dong, X.; Zheng, L.; Ma, F.; Yang, Y.; and Meng, D. 2018. Few-example object detection with model communication. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7): 1641–1654.

Du, S.; Tao, Y.; and Martinez, A. M. 2014. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15): E1454–E1462.

Ekman, P.; and Friesen, W. V. 1971. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2): 124–129.

Fabian Benitez-Quiroz, C.; Srinivasan, R.; and Martinez, A. M. 2016. EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 5562–5570.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the International Conference on Machine Learning*, 1126–1135.

Garcia, V.; and Bruna, J. 2018. Few-shot learning with graph neural networks. In *Proceedings of the International Conference on Learning Representations*.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*.

Guo, J.; Zhou, S.; Wu, J.; Wan, J.; Zhu, X.; Lei, Z.; and Li, S. Z. 2017. Multi-modality network with visual and geometrical information for micro emotion recognition. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, 814–819.

Guo, Y.; Codella, N. C.; Karlinsky, L.; Codella, J. V.; Smith, J. R.; Saenko, K.; Rosing, T.; and Feris, R. 2020. A broader study of cross-domain few-shot learning. In *Proceedings of the European Conference on Computer Vision*, 124–141.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 770–778.

Hu, Y.; Gripon, V.; and Pateux, S. 2021. Leveraging the feature distribution in transfer-based few-shot learning. In *Proceedings of the International Conference on Artificial Neural Networks*, 487–499.

Kingma, D. P.; and Ba, J. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*.

Lake, B. M.; Salakhutdinov, R.; and Tenenbaum, J. B. 2015. Human-level concept learning through probabilistic program induction. *Science*, 350(6266): 1332–1338.

Li, S.; Deng, W.; and Du, J. 2017. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2852–2861.

Li, W.; Xu, J.; Huo, J.; Wang, L.; Gao, Y.; and Luo, J. 2019. Distribution consistency based covariance metric networks for few-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 8642–8649.

Li, Y.; Zeng, J.; Shan, S.; and Chen, X. 2018. Occlusion aware facial expression recognition using CNN with attention mechanism. *IEEE Transactions on Image Processing*, 28(5): 2439–2450.

Liu, C.; Fu, Y.; Xu, C.; Yang, S.; Li, J.; Wang, C.; and Zhang, L. 2021. Learning a few-shot embedding model with contrastive Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 8635–8643.

Lu, J.; Gong, P.; Ye, J.; and Zhang, C. 2020. Learning from very few samples: A survey. *arXiv preprint arXiv:2009.02653*.

Lucey, P.; Cohn, J. F.; Kanade, T.; Saragih, J.; Ambadar, Z.; and Matthews, I. 2010. The extended Cohn-Kanade dataset

(CK+): A complete dataset for action unit and emotion-specified expression. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 94–101.

Luo, Z.; Zou, Y.; Hoffman, J.; and Fei-Fei, L. 2017. Label efficient learning of transferable representations across domains and tasks. In *Advances in Neural Information Processing Systems*, 165–177.

Pantic, M.; Valstar, M.; Rademaker, R.; and Maat, L. 2005. Web-based database for facial expression analysis. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, 317–321.

Patacchiola, M.; Turner, J.; Crowley, E. J.; O'Boyle, M.; and Storkey, A. 2020. Bayesian meta-learning for the few-shot setting via deep kernels. In *Advances in Neural Information Processing Systems*.

Ruan, D.; Yan, Y.; Chen, S.; Xue, J.; and Wang, H. 2020. Deep disturbance-disentangled learning for facial expression recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2833–2841.

Slimani, K.; Lekdioui, K.; Messoussi, R.; and Touahni, R. 2019. Compound facial expression recognition based on highway CNN. In *Proceedings of the New Challenges in Data Sciences: Acts of the Second Conference of the Moroccan Classification Society*, 1–7.

Snell, J.; Swersky, K.; and Zemel, R. S. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, 4077–4087.

Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H.; and Hospedales, T. M. 2018. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 1199–1208.

Tseng, H.-Y.; Lee, H.-Y.; Huang, J.-B.; and Yang, M.-H. 2020. Cross-domain few-shot classification via learned feature-wise transformation. In *Proceedings of the International Conference on Learning Representations*.

Vinyals, O.; Blundell, C.; Lillicrap, T.; Kavukcuoglu, K.; and Wierstra, D. 2016. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, 3630–3638.

Yang, S.; Liu, L.; and Xu, M. 2021. Free lunch for few-shot learning: Distribution calibration. In *Proceedings of the International Conference on Learning Representations*.

Yang, Z.; Wang, Y.; Chen, X.; Liu, J.; and Qiao, Y. 2020. Context-transformer: Tackling object confusion for few-shot detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 12653–12660.

Yao, H.; Zhang, C.; Wei, Y.; Jiang, M.; Wang, S.; Huang, J.; Chawla, N.; and Li, Z. 2020. Graph few-shot learning via knowledge transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 6656–6663.

Zhang, K.; Zhang, Z.; Li, Z.; and Qiao, Y. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10): 1499–1503.

Zhao, G.; Huang, X.; Taini, M.; Li, S. Z.; and PietikäInen, M. 2011. Facial expression recognition from near-infrared videos. *Image and Vision Computing*, 29(9): 607–619.

Zhao, Z.; Liu, Q.; and Zhou, F. 2021. Robust lightweight facial expression recognition network with label distribution training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 3510–3519.