https://doi.org/10.2352/EI.2026.38.7.IMAGE-267

Dual-Stream Feature Disentanglement Network for Single Domain Generalized Facial Expression Recognition

Ningyu Chen[1], Wenshui Lin[1], Chang Shu[2], Yan Yan [1,*]
[1] Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, School of Informatics, Xiamen University, China
[2] School of Communication and Information Engineering, University of Electronic Science and Technology of China, China

Abstract

Over the past few decades, facial expression recognition (FER) has been widely deployed in real-world applications. However, the collection conditions of existing datasets vary substantially, leading to significant domain shifts among datasets. Consequently, the performance of the most advanced FER methods will deteriorate in cross-domain scenarios. To address this issue, we propose a Dual-Stream Feature Disentanglement Network (DFD-Net) within the Single Domain Generalization (SDG) paradigm. DFD-Net employs the Expression Feature Extraction (EFE) module together with an attention block as the expression feature extraction branch, performing primary feature fusion and high-level feature selection. In parallel, the Expression-Irrelevant Feature Extraction (EIFE) module and the Expression-Irrelevant Feature Predictor (EIFP) constitute another branch. EIFE is pre-trained to capture the expression-irrelevant feature. EIFP passes the expression features through the Gradient Reversal Layer (GRL) and the Mutual Information Predictor (MIP) to compute and minimize the mutual information with the expression-irrelevant features. Extensive experiments on multiple benchmark datasets demonstrate that our method consistently outperforms existing state-of-the-art methods.

Introduction

Facial expressions constitute a fundamental modality of human social interaction, serving as a primary way for the transmission of affective information and the externalization of internal emotional states [1]. Humans can infer an individual's emotional state at a certain moment, such as happiness, sadness, anger, or surprise. In recent years, the rapid advancement of deep learning has resulted in a proliferation of deep neural architectures tailored for facial expression recognition (FER) [2, 3, 4]. These developments have facilitated the widespread deployment of automated expression analysis systems in a broad spectrum of applications, including human–computer interaction, health monitoring, affective computing, and social media surveillance [5, 6].

Cross-domain discrepancies constitute a primary impediment to the generalization capacity of FER models. Such discrepancies arise because datasets are typically collected under divergent environmental conditions (e.g., illumination, camera devices). Moreover, facial data are sourced from different individuals, and expression-irrelevant factors (e.g., gender, race, age, identity) among individuals vary across datasets. As a result, the expression features extracted by FER models in different datasets tend to be only partially aligned, leading to substantial performance degradation even of advanced FER models in cross-domain scenarios [7].

In recent years, numerous Cross-Domain Facial Expression Recognition (CD-FER) methods have been proposed. Most of these methods have primarily concentrated on Domain Adaptation (DA) [7, 8], where models are trained on images of the source domain annotated with expression labels, while unlabeled target domain images are utilized to mitigate domain shift. However, in practical scenarios, collecting facial expression images from the target domain is often infeasible. This highlights the need to develop CD-FER methods that do not rely on target domain data.

In this paper, we adopt the experimental setting of Single Domain Generalization (SDG) to analyze the CD-FER model. Compared with DA, SDG does not require unlabeled samples from the target domain to participate in model training. This setting better reflects real-world deployment scenarios. A well-performing SDG-FER model should demonstrate excellent zero-shot capability on various expression datasets.

We observe that due to domain bias, state-of-the-art FER models often fail to generalize effectively to unseen target domain samples. For two facial expression samples originating from different datasets, their expression-irrelevant factors (e.g., gender, race, age, identity, pose, illumination) can exhibit substantial differences. When an FER model is trained exclusively on the source domain, the lack of sufficient target domain samples frequently leads to overfitting to the source domain distribution. Moreover, in the unseen target domain, the joint distribution of expression-related and expression-irrelevant features differs from that of the source domain. Such distributional discrepancies hinder the model's ability to generalize to unseen target domains during feature processing, thus degrading its performance.

In brief, the overall pipeline of our proposed DFD-Net is a dual-stream architecture designed to decouple expression-relevant features from irrelevant attributes. Specifically, the Expression Feature Extraction (EFE) module and the attention block together form the expression feature extraction branch, which leverages multi-level feature fusion and an attention mechanism to obtain more robust expression features. The Expression-Irrelevant Feature Extraction (EIFE) module and the Expression-Irrelevant Feature Predictor (EIFP) constitute another branch. The EIFE module is pre-trained on a large-scale facial dataset annotated with expression-irrelevant factors, enabling it to effectively extract expression-irrelevant features. The EIFP incorpo-

*Corresponding Author

rates Gradient Reversal Layer (GRL) [9] and Mutual Information Predictor (MIP), the expression features are passed through the GRL and then fed into the MIP. MIP computes and minimizes the mutual information between prediction results and expression-irrelevant features, thus eliminating expression-irrelevant information from the model during backpropagation.

In summary, the contributions of this paper are as follows:

- We propose DFD-Net for single domain generalized facial expression recognition. DFD-Net leverages large-scale facial datasets annotated with expression-irrelevant information for pretraining, enabling it to effectively extract expression-irrelevant features from facial images, which are subsequently utilized for mutual information minimization.
- We design a dual-branch architecture in DFD-Net, where EIFE is dedicated to extracting expression-irrelevant features, while EFE focuses on capturing facial expression features. For EFE, an attention block is appended after the backbone to further refine feature selection by integrating shallow and deep features. In addition, EIFP enables the model to eliminate expression-irrelevant features during backpropagation.
- To demonstrate the effectiveness of our method, we evaluated it on multiple popular facial expression datasets. The results showed that our method significantly outperformed the SOTA expression recognition method.

Related Work

Facial Expression Recognition. Current facial expression recognition (FER) methods can achieve good performance on a single dataset. Zhang et al. [10] propose a weakly supervised local–global relation network for FER. Ruan et al. [11] propose a method that utilizes the attention mechanism and adversarial learning to separate the interfering factors in FER. However, when domain bias occurs between the training data and the test data, these FER methods often experience performance degradation. This paper aims to train the model using a single source domain sample and ensure that the trained model performs well on several unseen target datasets.

Cross-Domain Facial Expression Recognition. Several CD-FER methods have been proposed under the DA paradigm. Chen et al. [8] benchmark the CD-FER method under the DA setting and propose an adversarial graph learning method that combines global and local features of the face. Li et al. [12] propose a method based on mutual information and discriminative metric learning techniques to bridge the semantic differences between the source domain and the target domain. However, these methods rely on images from the unlabeled target domain during training. In practical deployment scenarios, even unlabeled target domain images are often difficult to obtain. Our method differs from these methods: during the training process of the proposed DFD-Net, a single FER dataset is used as the training set, and the model is evaluated on multiple unseen FER datasets. Consequently, existing CD-FER methods developed under DA settings are not applicable to our task.

Problem Definition

Our objective is to address the limited generalization capability in CD-FER caused by the unavailability of target domain training images. We aim to train the model on a single source domain dataset while ensuring that the learned model performs well on multiple unseen target domain. These target domain samples exhibit domain discrepancies with respect to the training set and remain inaccessible during training, which closely aligns with the deployment scenario of FER models in real-world applications.

Specifically, in our experimental setting, the FER model is trained on a source domain training set $\mathcal{D}^s_{\text{train}} = \{(x^s_i, y^s_i)\}_{i=1}^{N_s}$, where x^s_i denotes the $i-th$ facial expression image from the source domain, $y^s_i \in Y = \{1, \ldots, L\}$ is its corresponding expression label, N_s is the number of training samples in the source domain, and L is the number of expression classes. Following existing studies, we adopt the seven basic expressions. To better simulate real-world deployment, the model is evaluated on the target domain test set $\mathcal{D}^t_{\text{test}} = \{(x^t_i, y^t_i)\}_{i=1}^{N_t}$, where x^t_i denotes the $i-th$ facial expression image from the target domain, y^t_i is the corresponding label, and N_t is the number of test samples in the target domain.

Our setting is different from traditional FER and DA-FER. In traditional FER, the model is trained and tested on datasets drawn from the same distribution, where $\mathcal{D}^s_{\text{train}}$ and $\mathcal{D}^s_{\text{test}}$ belong to the same domain. In DA-FER, the training set includes source and target data, where $\mathcal{D}_{\text{train}} = \mathcal{D}^s_{\text{train}} \cup \mathcal{D}^t_{\text{train}}$, unlabeled samples from the target domain training set $\mathcal{D}^t_{\text{train}}$ are utilized to fine-tune the model, and the evaluation is carried out on the target domain test set $\mathcal{D}^t_{\text{test}}$ with domain shift. In the SDG setting, only the source domain training set $\mathcal{D}^s_{\text{train}}$ is used for training. The trained model is then deployed directly on target domain test sets $\mathcal{D}^t_{\text{test}}$ with different distributions, where the target domain data is completely unavailable during training. This setting better reflects the deployment of real-world models and poses greater challenges.

Methodology

Overview

To address the aforementioned problem, we design the Dual-Stream Feature Disentanglement Network (DFD-Net), as illustrated in Fig. 1. DFD-Net consists of two training stages. In the first stage, the Expression-Irrelevant Feature Extraction (EIFE) module is trained on large-scale facial datasets annotated with distraction factors (RAF-DB [13] and Multi-PIE [14]), to extract expression-irrelevant attributes such as gender, race, age, identity, pose, and illumination. These extracted expression-irrelevant features are used later for mutual information minimization in the second stage. In the second stage, we train the Expression Feature Extraction (EFE) module along with the subsequent attention block. The EFE fuses shallow and deep features and dynamically adjusts their weights to compute a weighted sum as input to the attention block. The attention block then generates attention weights, which are multiplied by deep features to obtain attention-refined features. These attention features are fed into the classification head for expression prediction. Simultaneously, the features are also fed into the Expression-Irrelevant Feature Predictor (EIFP), where they pass through the Gradient Reversal Layer (GRL) [9] and subsequently enter the Mutual Information Predictor (MIP) to predict mutual information. The predicted results are then compared with the outputs of EIFE to compute and minimize their mutual information. This process enables the disentanglement of expression-irrelevant factors from expression-related features during backpropagation.

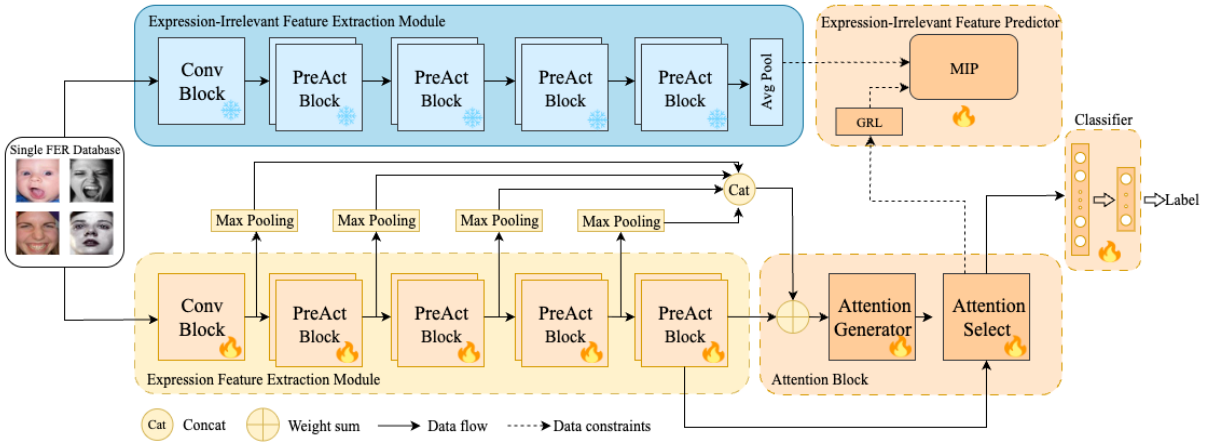


Figure 1. *Overview of DFD-Net.*

Expression-Irrelevant Feature Extraction Module

To extract expression-irrelevant features from facial images, we employ large-scale facial datasets annotated with expression-irrelevant factors to train EIFE. By training EIFE, expression-irrelevant information can be effectively captured within facial images. This enables the use of EIFP in the second training stage to obtain domain-invariant expression features.

Specifically, the RAF-DB dataset provides gender, race, and age annotations [13], while the Multi-PIE dataset provides identity, pose, and illumination annotations [14]. We employ different backbones for the two datasets and design separate fully connected layers for classification, thereby enabling the capture of expression-irrelevant information from facial images.

Given the facial training dataset $\mathcal{D}_n = \{(x_i, z_i)\}_{i=1}^{N}$, where x_i is the $i-th$ training image, and $z_i \in \mathbb{R}^M$ is the vector of labels corresponding to M expression-irrelevant attributes, M represents the number of expression-irrelevant factors. The loss of classification for the $j-th$ expression-irrelevant factor can be formulated as :

$$\mathcal{L}_{\text{cls}}^{j} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left((\mathbf{w}_{z_i^j}^{j})^T\cdot\mathbf{f}_i\right)}{\sum_{c=1}^{C_j}\exp\left((\mathbf{w}_c^j)^T\cdot\mathbf{f}_i\right)}, \tag{1}$$

where $\mathbf{f}_i$ denotes the feature representation of sample x_i obtained from the EIFE and the task-specific branch, $\mathbf{w}_c^j$ is the classification weight vector corresponding to the $c-th$ category of the $j-th$ expression-irrelevant factor, z_i^j is the ground-truth label of the $i-th$ sample for the $j-th$ expression-irrelevant factor, C_j represents the number of classes associated with the $j-th$ expression-irrelevant factor. Accordingly, the overall optimization objective of EIFE is defined as follows:

$$\min_{\theta_s,\{\theta_j\}_{j=1}^{M}}\sum_{j=1}^{M}\mathcal{L}_{\text{cls}}^{j}, \tag{2}$$

where θ_s represents the shared parameters, θ_j corresponds to the prediction branch of the $j-th$ interference factor. Through this training strategy, we obtain the EIFE model capable of extracting expression-irrelevant features, which is then utilized in the second training stage.

Expression Feature Extraction Module and Attention Block

In the second stage, we train the EFE together with the attention block. Specifically, we adopt PreAct ResNet-18 [15] as the EFE backbone to extract facial expression features. The multi-level fused features and the deep features are fed into the attention block, where two learnable parameters dynamically weight their contributions to compute the attention mask. Finally, the output of the attention block is fed into both the classifier and the EIFP, both conducting expression prediction and mutual information minimization calculations respectively.

Given the single FER training dataset $\mathcal{D}_{\text{train}}^{s} = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$, where x_i^s is the $i-th$ training image and y_i^s is the expression label corresponding to x_i^s. For a facial expression image x_i^s from the training set, it is passed sequentially through each layer of the EFE. After each layer, max-pooling is applied to reduce the feature dimensions to match those of the final layer. The resulting features are then concatenated along the channel dimension, thereby achieving the fusion of shallow and deep features. The feature fusion process can be formulated as:

$$\mathbf{f}_{\text{fused}}^{e} = \left[\mathbf{f}_1^e\ ,\ \cdots\ ,\ \mathbf{f}_{L-1}^e\right], \tag{3}$$

where $\mathbf{f}_j^e$ denotes the output of the $j-th$ residual block in PreAct ResNet-18 after max-pooling, and $\mathbf{f}_{\text{fused}}^{e}$ represents the fused feature map. After obtaining the multi-level fused features, we introduce two learnable parameters to dynamically adjust the relative contributions of the fused features and the deep features in the subsequent classification, these weighted features are then used as the input to the attention block.

The attention block structure is illustrated in Fig. 2. Specifically, the **Mask** generated by the attention generator within the attention block can be expressed as:

$$\textbf{Mask} = g(\delta_1\mathbf{f}_{\text{fused}}^{e} + \delta_2\mathbf{f}_L^e), \tag{4}$$

where δ_1 and δ_2 are learnable parameters, and $g(\cdot)$ denotes the attention generator. The attention generator employs a series of convolution operations followed by a Sigmoid activation to constrain the values of the attention mask within the range $(0,1)$, thus assigning an attention weight to each individual element.

After obtaining the attention mask, it is multiplied element-wise with the final-layer output of the EFE model $\mathbf{f}_L^e$ to generate

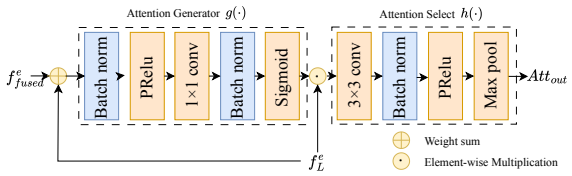


Figure 2. *Overview of Attention Block.*

the attention-selected features, which can be formulated as:

$$\mathbf{Att}_{out} = h(\mathbf{Mask} \odot \mathbf{f}^e_L), \tag{5}$$

where $\odot$ denotes element-wise multiplication, and $h(\cdot)$ represents the attention selection process, which consists of a series of convolution operations, batch normalization, and PReLU activation, followed by a max-pooling operation. The features computed by the attention block are then fed into the classifier and the EIFP for subsequent computations.

After the output of the attention block, a classification head is attached to predict the expression category. The classification loss is defined as:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\left((\mathbf{w}_{y_i})^T \cdot \mathbf{Att}^i_{out}\right)}{\sum_{c=1}^{C} \exp\left((\mathbf{w}_c)^T \cdot \mathbf{Att}^i_{out}\right)}, \tag{6}$$

where $\mathbf{Att}^i_{out}$ denotes the feature representation of the $i-th$ sample obtained from the attention block, $\mathbf{w}_c$ is the weight vector in the classification head corresponding to the $c-th$ expression class, y_i is the ground-truth expression label of the $i-th$ sample, and C is the total number of expression categories.

Expression-Irrelevant Feature Predictor

In the first stage, we train EIFE using large-scale facial datasets annotated with expression-irrelevant factors. In the second stage, we design the EIFP, which includes GRL and MIP, enabling the removal of expression-irrelevant factors from facial features within the EFE and the attention block during backpropagation.

Suppose that the EIFE produces two expression-irrelevant features $\mathbf{d}_1$ and $\mathbf{d}_2$, we design two predictors MI_1 and MI_2, whose inputs are the outputs of $\mathbf{Att}_{out}$ after passing through the GRL:

$$\hat{\mathbf{d}}_1 = \text{MI}_1\big(\text{GRL}(\mathbf{Att}_{out})\big), \quad \hat{\mathbf{d}}_2 = \text{MI}_2\big(\text{GRL}(\mathbf{Att}_{out})\big), \tag{7}$$

where $\hat{\mathbf{d}}_1$ and $\hat{\mathbf{d}}_2$ represent the mutual information prediction results of MIP. For the EIFP, the loss function is defined as:

$$\mathcal{L}_{\text{MI}} = \mathcal{L}_{\text{reg}}\big(\hat{\mathbf{d}}_1, \mathbf{d}_1\big) + \mathcal{L}_{\text{reg}}\big(\hat{\mathbf{d}}_2, \mathbf{d}_2\big), \tag{8}$$

where $\mathcal{L}_{\text{reg}}$ adopts the smooth L_1 loss. Although the loss function adopts the smooth L_1 form, the GRL-based adversarial training forces the EFE and the attention block to discard features that are predictive of expression-irrelevant factors, thus indirectly minimizing the mutual information between expression features and expression-irrelevant features.

Specifically, during forward propagation, the GRL acts as an identity mapping: $\text{GRL}(\mathbf{Att}_{out}) = \mathbf{Att}_{out}$, during backpropagation, it reverses the gradients and scales them by a factor of $-\lambda$:

$$\frac{\partial \mathcal{L}_{\text{MI}}}{\partial \mathbf{Att}_{out}} = -\lambda \cdot \frac{\partial \mathcal{L}_{\text{MI}}}{\partial \text{GRL}(\mathbf{Att}_{out})}, \tag{9}$$

where MIP try to recover the distracting features $\mathbf{d}_1$ and $\mathbf{d}_2$ from $\mathbf{Att}_{out}$. However, due to gradient reversal, the EFE and attention block is updated with reversed gradients, forcing $\mathbf{Att}_{out}$ to reduce its correlation with $\mathbf{d}_1$ and $\mathbf{d}_2$.

Joint Loss Function

The joint loss function for the model is given as:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \gamma \mathcal{L}_{\text{MI}}, \tag{10}$$

where γ is a balancing coefficient used to dynamically adjust the relative weight between expression classification and adversarial regularization. In implementation, is scheduled according to the training progress $p \in [0,1]$:

$$\gamma(p) = \frac{2}{1+\exp(-10p)} - 1. \tag{11}$$

Instead of a fixed weight, this dynamic strategy stabilizes adversarial training. Classification learning is emphasized in the early steps of training, while the effect of adversarial regularization is gradually strengthened in the later steps.

Experiment

In this section, we conduct experiments on the proposed DFD-Net. First, we introduce our experimental setup, including the datasets used and the details of implementation of DFD-Net. Next, we performed ablation studies to demonstrate the effectiveness of its components. Finally, we compare DFD-Net with several state-of-the-art methods.

Experimental Settings

Datasets We employ two datasets for training in the first stage. In the second stage, we adopt the single domain generalization setting, where one dataset is used as the source domain for training, while the test sets of the remaining datasets are utilized for evaluation. Classification accuracy is used as the performance metric. **RAF-DB** [13] is a dataset collected from the Internet, each image is annotated not only with one of the seven basic expression labels but also with age, gender, and ethnicity attributes. Therefore, in this work, RAF-DB is utilized for model training at first stage. RAF-DB including 12,271 images for training and 3,068 images for testing. **SFEW 2.0** [16] is an in-the-wild dataset, collected from various movies. All images are annotated with one of the seven basic expressions. SFEW 2.0 contains 958 training samples, 436 validation samples, and 372 test samples. **MMA** is a large-scale facial expression database primarily composed of facial images of individuals of European and American origin. It is categorized into seven basic expression classes and comprises 92,968 training samples, 17,356 validation samples, and 17,356 testing samples. **FERPlus** is an in-the-wild dataset, which serves as an extended version of FER2013 [17]. It contains 28,709 training images and 3,589 test images, annotated with cleaner expression labels.

Implementation Details For the training in the first stage, we train two separate EIFEs on RAF-DB and Multi-PIE. The EIFE

Table 1: Ablation Studies for DFD-Net. Use RAF-DB as the source domain.

Att	EIFP	FERPlus	SFEW 2.0	MMA	Mean
		58.05	42.76	42.61	47.81
✓		69.30	48.17	47.94	55.14
✓	✓	**71.79**	**49.31**	**50.34**	**57.15**

trained on RAF-DB is used to capture gender, race, and age feature from facial images, while the EIFE trained on Multi-PIE is used to capture identity, pose, and illumination feature.

For the training in the second stage, we train the EFE, the attention block, and the EIFP on a single FER dataset. Each input facial image is randomly cropped and resized to 90×90, with the scaling factor sampled from the range [0.8, 1.0]. To improve robustness to illumination and color variations, color jitter is applied to brightness, contrast, saturation, and hue, with a maximum variation of 0.4. A random horizontal flip with a probability of 0.5 is applied. After these augmentations, the images are converted into tensors and normalized using the ImageNet mean and standard deviation [18]. We use PreAct ResNet-18 [15], which is pretrained on AffectNet [19], as the backbone and optimize it with AdamW [20], with $\beta_1 = 0.5$ and $\beta_2 = 0.999$, setting the initial learning rate to 1e-6. For the attention block, EIFP, and the classifier, the initial learning rate is set to 5e-4. A cosine annealing learning rate schedule is applied to all components, with T_{max} set to 100. The model is trained for 60 epochs on a single NVIDIA RTX 3090 GPU with a batch size of 128.

Ablation Studies

To investigate the effectiveness of each proposed module in our method, we conduct ablation studies on DFD-Net. The results in Table 1 show that without any additional modules, the FER model trained on RAF-DB cannot generalize well to other datasets with domain gaps, achieving only 47.81% on average. With the introduction of the attention block, which fuses deep and shallow features, the performance on unseen target datasets improves substantially, surpassing the baseline by a large margin. This shows that attention effectively enhances the ability of the network to capture expression-relevant information. However, relying solely on attention is not sufficient to remove disturbance features, which still limits the generalizability of the model. To address this issue, we further introduce the EIFE and EIFP. By leveraging the pretrained EIFE to extract expression-irrelevant features, we further employ EIFP to perform gradient reversal and mutual information minimization, expression-irrelevant factors can be effectively eliminated during backpropagation. As a result, performance improves consistently across all target datasets, and the model achieves the best overall accuracy of 57.15%. From the results in Table 1 , it can be observed that both modules contribute to the success of DFD-Net, and combining them together achieves the strongest cross-domain generalization.

Comparison with State-of-the-Art Methods

We compare the proposed DFD-Net with existing facial expression recognition methods. Specifically, we use one of the four datasets as the training set and evaluate the model on the remaining three target domains, where facial images from the target domains are inaccessible during training. The test results are presented in Table 2. DFD-Net consistently outperforms the state-of-the-art FER methods under the cross-domain setting. For example, when trained on RAF-DB, DFD-Net achieves 71.79%, 49.31%, and 50.34% on FERPlus, SFEW 2.0, and MMA, with a mean accuracy of 57.15%, which significantly surpasses the second-best method RUL (mean 47.30%). Similarly, when trained on FERPlus, DFD-Net reaches a mean accuracy of 59.01%, showing clear improvements over RUL, EAC, and OFER. Moreover, under other training settings (SFEW 2.0 or MMA as the source domain), DFD-Net still achieves the best or comparable results across multiple target datasets. These results demonstrate the superior generalization ability of our model and verify its effectiveness in handling domain shifts, highlighting the advantage of explicitly modeling disturbance features and enhancing expression-related representations.

Table 2: Comparison of different methods.

Method	FERPlus	SFEW	MMA	Mean
RUL (NeurIPS2021) [21]	57.89	46.91	37.11	47.30
EAC (ECCV2022) [22]	54.38	43.39	37.27	45.01
OFER (ICCV2023) [23]	53.90	43.88	36.43	44.74
DFD-Net	**71.79**	**49.31**	**50.34**	**57.15**
Method	RAF-DB	SFEW	MMA	Mean
RUL (NeurIPS2021) [21]	51.89	45.90	58.00	51.93
EAC (ECCV2022) [22]	58.62	45.79	**59.87**	54.76
OFER (ICCV2023) [23]	55.67	44.52	59.37	53.19
DFD-Net	**71.54**	**48.62**	56.86	**59.01**
Method	RAF-DB	FERPlus	MMA	Mean
RUL (NeurIPS2021) [21]	46.35	36.02	22.06	34.81
EAC (ECCV2022) [22]	47.29	33.66	22.26	34.40
OFER (ICCV2023) [23]	46.33	34.78	21.88	34.33
DFD-Net	**64.86**	**61.11**	**39.05**	**55.01**
Method	RAF-DB	FERPlus	SFEW	Mean
RUL (NeurIPS2021) [21]	71.94	69.05	39.21	60.07
EAC (ECCV2022) [22]	**74.32**	71.85	42.87	63.01
OFER (ICCV2023) [23]	72.85	69.69	41.49	61.34
DFD-Net	74.28	**74.63**	**51.61**	**66.84**

Conclusion

In this paper, we propose DFD-Net to address the problem of SDG-FER. DFD-Net adopts a dual-branch architecture comprising two branches, EIFE and EFE, which are designed to extract expression-irrelevant features and expression-related features. For EFE, multi-level feature fusion and an attention block are employed for further feature extraction. To remove the expression-irrelevant feature, we design the EIFP, which includes the MIP and GRL. MIP minimizes the mutual information between expression and expression-irrelevant features, while GRL enforces the features extracted by the EFE and attention blocks to contain fewer expression-irrelevant features during backpropagation. Experimental results demonstrate that DFD-Net compared with baseline can be better generalized to the unseen target domain. Furthermore, we conduct comparisons with state-of-the-art FER methods, where DFD-Net consistently achieves better performance.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62372388, Grant U21A20514 and Grant 62571466, the Major Science and Technology Plan Project on the Future Industry Fields of Xiamen City under Grant 3502Z20241029 and Grant 3502Z20241027, the Fundamental Research Funds for the Central Universities under Grant 20720240076 and Grant ZYGX2021J004.

References

[1] C. Darwin and S. F. Darwin, *The expression of the emotions in man and animals*. John Murray London, 1872, vol. 3.

[2] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE transactions on affective computing*, vol. 13, no. 3, pp. 1195–1215, 2020.

[3] D. Ruan, R. Mo, Y. Yan, S. Chen, J.-H. Xue, and H. Wang, "Adaptive deep disturbance-disentangled learning for facial expression recognition," *International Journal of Computer Vision*, vol. 130, no. 2, pp. 455–477, 2022.

[4] R. Ni, B. Yang, X. Zhou, A. Cangelosi, and X. Liu, "Facial expression recognition through cross-modality attention fusion," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 15, no. 1, pp. 175–185, 2022.

[5] F. Zhang, T. Zhang, Q. Mao, L. Duan, and C. Xu, "Facial expression recognition in the wild: A cycle-consistent adversarial attention transfer approach," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 126–135.

[6] F. Zhang, T. Zhang, Q. Mao, and C. Xu, "Joint pose and expression modeling for facial expression recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3359–3368.

[7] Y. Gao, Y. Xie, Z. Z. Hu, T. Chen, and L. Lin, "Adaptive global-local representation learning and selection for cross-domain facial expression recognition," *IEEE Transactions on Multimedia*, vol. 26, pp. 6676–6688, 2024.

[8] T. Chen, T. Pu, H. Wu, Y. Xie, L. Liu, and L. Lin, "Cross-domain facial expression recognition: A unified evaluation benchmark and adversarial graph learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 12, pp. 9887–9903, 2021.

[9] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.

[10] H. Zhang, W. Su, J. Yu, and Z. Wang, "Weakly supervised local-global relation network for facial expression recognition," in *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*, 2021, pp. 1040–1046.

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

[12] Y. Li, Y. Gao, B. Chen, Z. Zhang, L. Zhu, and G. Lu, "Jdman: Joint discriminative and mutual adaptation networks for cross-domain facial expression recognition," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3312–3320.

[13] S. Li, W. Deng, and J. Du, "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*, 2017, pp. 2852–2861.

[14] G. Ralph, M. Iain, C. Jeffrey, K. Takeo, and B. Simon, "Multi-pie," *Image and vision computing*, vol. 28, no. 5, pp. 807–813, 2010.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European conference on computer vision*. Springer, 2016, pp. 630–645.

[16] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," in *2011 IEEE international conference on computer vision workshops (ICCV workshops)*. IEEE, 2011, pp. 2106–2112.

[17] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *Proceedings of the 18th ACM international conference on multimodal interaction*, 2016, pp. 279–283.

[18] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[19] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.

[20] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[21] Y. Zhang, C. Wang, and W. Deng, "Relative uncertainty learning for facial expression recognition," *Advances in Neural Information Processing Systems*, vol. 34, pp. 17 616–17 627, 2021.

[22] Y. Zhang, C. Wang, X. Ling, and W. Deng, "Learn from all: Erasing attention consistency for noisy label facial expression recognition," in *European Conference on Computer Vision*. Springer, 2022, pp. 418–434.

[23] I. Lee, E. Lee, and S. B. Yoo, "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*, 2023, pp. 1536–1546.

Author Biography

Ningyu Chen *is a master's student majoring in computer science at Xiamen University in China. His main research area is computer vision.*

Wenshui Lin *is an assistant professor at Xiamen University, China. He received his Ph.D. degree from Xiamen University, China, in 2007. His main research area is machine learning.*

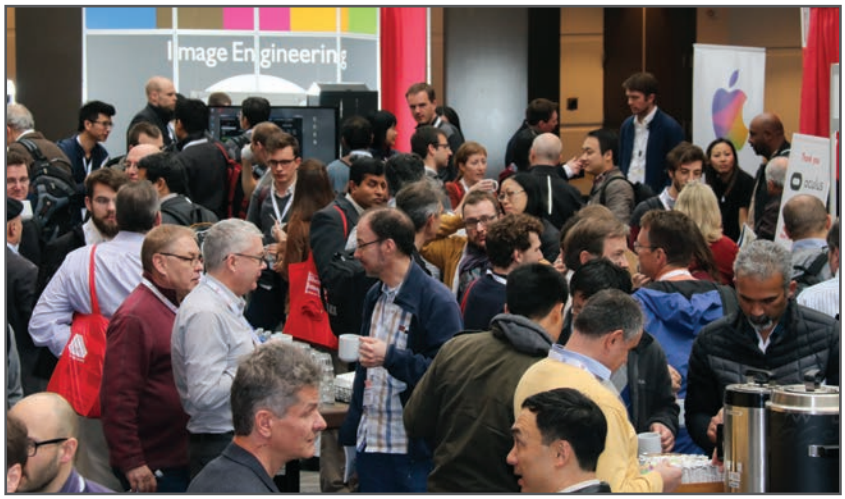
Chang Shu *is a lecturer at the University of Electronic Science and Technology of China. He received his Ph.D. degree from Tsinghua University, China, in 2011. His main research area is computer vision.*

Yan Yan *is a professor at Xiamen University, China. He received his Ph.D. degree from Tsinghua University, China, in 2009. His main research area is computer vision.*

JOIN US AT THE NEXT EI!

electronic IMAGING

Imaging across applications . . . Where industry and academia meet!



• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

ADVANCING IMAGING SCIENCE 75 1947 IS&T 2022 imaging.org