Cross-Domain Facial Expression Recognition via Disentangling Identity Representation

Tong Liu[1], **Jing Li**[1], **Jia Wu**[2], **Lefei Zhang**[1], **Shanshan Zhao**[3], **Jun Chang**[1] and **Jun Wan**[4]
[1]Wuhan University, China
[2]Macquarie University, Sydney
[3]JD Explore Academy, China
[4]Zhongnan University of Economics and Law, China

Abstract

Most existing cross-domain facial expression recognition (FER) works require target domain data to assist the model in analyzing distribution shifts to overcome negative effects. However, it is often hard to obtain expression images of the target domain in practical applications. Moreover, existing methods suffer from the interference of identity information, thus limiting the discriminative ability of the expression features. We exploit the idea of domain generalization (DG) and propose a representation disentanglement model to address the above problems. Specifically, we learn three independent potential subspaces corresponding to the domain, expression, and identity information from facial images. Meanwhile, the extracted expression and identity features are recovered as Fourier phase information reconstructed images, thereby ensuring that the high-level semantics of images remain unchanged after disentangling the domain information. Our proposed method can disentangle expression features from expression-irrelevant ones (i.e., identity and domain features). Therefore, the learned expression features exhibit sufficient domain invariance and discriminative ability. We conduct experiments with different settings on multiple benchmark datasets, and the results show that our method achieves superior performance compared with state-of-the-art methods.

1 Introduction

Due to the difference in photographic environments (e.g., in-the-lab or in-the-wild) and collected subjects, there are obvious domain shifts across different facial expression datasets [Li and Deng, 2020]. The current well-performing FER methods may achieve satisfactory performance in intra-dataset protocols, but their performance drops dramatically in inter-dataset settings [Recht *et al.*, 2019; Zhang *et al.*, 2021]. Recently, a series of cross-domain FER algorithms [Chen *et al.*, 2021; Li *et al.*, 2022; Li *et al.*, 2021] have been widely developed to address this problem. They mainly focus on exploiting the idea of domain adaptation (DA) to collect data from each possible target domain and train the model with each source-target pair [Kang *et al.*, 2019; Huang *et al.*, 2021b]. However, in practice, there may not be any target domain data available during the training of expression recognition for the model to analyze the distribution shift to overcome the negative effects.

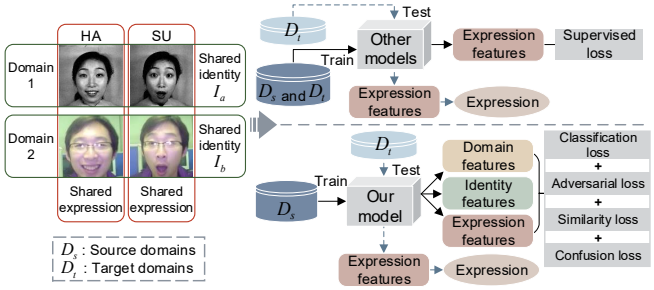


Figure 1: The high-level idea of our method. Other cross-domain FER models do not consider the influence of identity on cross-domain FER and require the target domain in the training process. Our method adopts the idea of DG and introduces feature disentanglement to disentangle the identity information from the domain and expression representation.

Recently, the DG methods have been studied extensively, which aim to generalize the knowledge extracted from multiple source domains to an unseen target domain that is not accessible during training [Gan *et al.*, 2016; Tzeng *et al.*, 2017]. Existing DG methods attempt to learn domain-agnostic features via employing various strategies such as adversarial feature learning [Li *et al.*, 2018], domain adversarial image generation [Zhou *et al.*, 2020], domain randomization [Huang *et al.*, 2021a], or style mixing [Zhou *et al.*, 2021b]. However, the domain-invariant features extracted by these methods may contain identity information, which makes the discrimination of the learned expression features limited. Eventually, the generalization of cross-domain FER is weakened. In fact, the identity information here is often the focus of face recognition or face verification work. In addition, the expressions to be recognized in real-world applications often come from unknown domains and subjects. It is necessary to eliminate the reliance on domain and identity information in the expression recognition process. As shown in Figure 1, cross-domain FER is interfered by the domain (different datasets) and identity information. Existing DG methods can hardly remove the

interference of identity information when learning domain-invariant features. Therefore, the discriminative capability of their extracted expression features is limited.

To solve the above issue, we extract more discriminative expression features via disentangling domain, expression, and identity information from the aspect of feature disentanglement [Zhou *et al.*, 2021a]. Thus the generalization performance of cross-domain FER is improved. Related studies [Muandet *et al.*, 2013; Zhang *et al.*, 2022] show that the learned representations reflect the intrinsic category semantics and have strong invariance to cross-domain variation, which is of great importance to the robustness and generalization of deep learning. Li et al. [Li *et al.*, 2022] also demonstrate that the learned representation is separable and domain-invariant, which is effective for cross-domain FER. Therefore, we handle cross-domain FER via learning domain-invariant, separable, and discriminative representations. As shown in Figure 1, we propose a representation disentanglement framework for extracting domain, expression, and identity features, respectively. Disentangling identity features from domain and expression features can enable our model to extract discriminative expression features in unseen domains. Moreover, to enforce the disentanglement of the above information and reduce the coupling between features, we provide additional supervision, which renders the disentangled expression features to be invariant to interference factors. Eventually, cross-domain FER can be better performed.

Additionally, we know from visual psychophysics that changes in the amplitude of the image Fourier transform can significantly alter the appearance but not the interpretation of the image. The high-level semantic information in images is associated with the phase of its Fourier transform [Hansen and Hess, 2007; Yang *et al.*, 2020]. Inspired by this, to ensure that the high-level semantic information of the facial image remains unchanged after disentangling domain information, we recover the encoded expression and identity features as images reconstructed from Fourier phase information. Accordingly, we construct the generator with an encoder-decoder structure that acts as an image semantic changer. It aims to output the fake Fourier phase information reconstructed image that can fool the phase information discriminator. Besides, to smooth the transformation of expressions, we introduce the identity discriminator to control the distribution of identity features. With the additional expression classifier, the decoder can strive for the generated image to have the same semantics as the real Fourier phase information reconstructed image. The above operations facilitate the disentangling of identity features and improve the discriminative ability of expression features. Meanwhile, they can ensure that the extracted expression features have good domain invariance.

The contributions of our work can be summarized as follows. (1) We propose a representation disentanglement network for domain-generalized FER, which can recognize facial expressions in unseen domains during inference. (2) We disentangle the identity features from both domain and expression features. Furthermore, the identity and expression features are recovered as the Fourier phase information reconstructed images. As a result, the discrimination and domain-invariance of expression features are all enhanced, which improves the performance of cross-domain FER. (3) With well-designed supervised learning strategies, the network modules are able to joint learning efficiently. Extensive experiments are conducted on different datasets. And the analytical results demonstrate the effectiveness and superiority of our method.

2 Related Work

2.1 Cross-Domain Facial Expression Recognition

To address the prevalent domain discrepancies among different FER datasets, some cross-domain FER algorithms have been recently proposed. For instance, Zhu et al. [Zhu *et al.*, 2016] propose a discriminative feature adaptive method to learn a feature space to represent facial images from different domains. Li et al. [Li and Deng, 2018] present a Deep Emotransfer Network to reduce the bias among the datasets. Chen et al. [Chen *et al.*, 2021] integrate graph propagation with adversarial learning mechanisms to learn domain-invariant holistic-local features for cross-domain FER. Li et al. [Li and Deng, 2020] propose a deep Emotion-Conditional Adaption Network to learn domain-invariant representations. However, the above methods require depending on the target domain data (without labels) to analyze the domain distribution shift. Li et al. [Li *et al.*, 2022] present a Deep Margin-Sensitive representation learning framework to extract multi-level features. But it depends highly on the supervision information on the target domain data since it needs to generate accurate pseudo-target labels. DLP-CNN [Li *et al.*, 2017b] aims to enhance feature discrimination by maximizing the inter-class dispersion while preserving local closeness. Zavarez et al. [Zavarez *et al.*, 2017] analyze the performance impact of fine-tuning with cross-database methods. Ji et al. [Ji *et al.*, 2019] propose an intra- and inter-class feature fusion network for FER across datasets. However, these approaches do not consider the influence of identity information on expression representation, thus limiting the discriminative power of expression features.

Different from the above approaches, we focus on the cross-domain FER when the target domain data is inaccessible, which is a more difficult case. We eliminate the dependence on identity during domain-invariant expression feature learning to improve the performance of cross-domain FER.

2.2 Domain Generalization

To overcome the problems of domain shift and lack of target domain data, DG is introduced. It aims to use the data from single or multiple related but different source domains to train the model that can generalize well to unseen target domains. It is evident that the DG model must depend only on the source domain to learn the domain-invariant representation. DG methods have been applied to many tasks, such as object recognition [Li *et al.*, 2018], image segmentation [Huang *et al.*, 2021a], and face anti-spoofing [Wang *et al.*, 2022]. Researchers categorize existing DG methods into different groups based on design motivations [Zhou *et al.*, 2021a], e.g., Domain Alignment [Shao *et al.*, 2019], Ensemble Learning [He *et al.*, 2016], Learning Disentangled Representations,

etc. Here we only briefly review the DG methods relevant to Learning Disentangled Representations. For example, Chen et al. [Chen *et al.*, 2016] utilize information-maximizing generative adversarial networks to learn disentangled representations. Chattopadhyay et al. [Chattopadhyay *et al.*, 2020] suggest learning domain-specific masks and encourage masks to learn a balance of domain-invariant and domain-specific features. Piratla et al. [Piratla *et al.*, 2020] propose a CSD that jointly learns public and domain-specific components. Since the facial expression is closely entangled with identity information, it is difficult for the above DG methods to disentangle them when learning the domain-invariant representation. Identity and expression representations are not disentangled, which limits the discriminative ability and generalization performance of expression features.

In contrast to the above DG methods, we analyze and consider the effect of identity information for the cross-domain FER task. Domain labels and generated Fourier phase information images are leveraged to disentangle features, which facilitates the model to learn domain-invariant, separable, and discriminative expression features.

3 Methodology

Forcing the entire model or features to be domain-invariant during cross-domain FER is challenging. We relax this constraint by allowing some parts to be domain-specific, i.e., learning disentangled representation. Intuitively, we learn three independent potential subspaces of the domain, expression, and identity information from facial images. In addition, the adversarial way is utilized to assist the model in learning domain-invariant and discriminative expression features.

3.1 Overview

Given a training set $\mathcal{D}_s = \{\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_S\}$ from multiple source domains, our goal is to learn a domain-agnostic FER model that is expected to perform well on the disjoint and unseen target domains $\mathcal{D}_t$. There are N_j labeled facial expression images $\{(x_i^j, y_i^j)\}_{i=1}^{N_j}$ in the j-th source domain $\mathcal{D}_j$, where x_i^j and $y_i^j \in \{0, 1, \cdots, K\}$ ($K = 6$ means the number of expression categories is 7) represent the input images and expression labels, respectively. The facial expressions in $\mathcal{D}_t$ will be recognized during the inference stage.

As shown in Figure 2, there are three encoders in our proposed network architecture for processing the input facial image: domain encoder E_d, expression encoder E_e, and identity encoder E_{id}. The domain encoder E_d is designed to extract the domain representation f_d from the input image, and the classifier C_d performs the domain classification task based on this representation. The expression representation f_e is extracted by the expression encoder E_e, followed by the expression classifier C_e for performing FER. The identity encoder E_{id} extracts the facial identity representation f_{id} from the images, while the decoder De is used to reconstruct the guarantee. The image still retains high-level semantics after disentangling the domain information, which can ensure further disentanglement of expression and identity representations. Inspired by the semantic preservation property of the Fourier phase component [Oppenheim and Lim, 1981; Xu *et al.*, 2021], the phase-only reconstructed images ($\hat{X}$) and phase information discriminator D_p are taken to guarantee the high-level semantics invariance of the images (X_p) reconstructed by the decoder De. The additional identity discriminator D_{id} controls the distribution of identity features and facilitates the smooth transformation of expressions. Once the joint learning of the above modules is completed, the cross-domain FER task can be accomplished simply through inference with E_e and C_e.

3.2 Learning Image Domain Representation

To address the domain-generalized FER, we suggest learning image domain representation from expression images. As shown in Figure 2, the learning of the domain representation is implemented by the E_d and C_d. Where the E_d is expected to describe information about background, illumination, resolution, etc. And the C_d recognizes the image domain j ($j \in \{1, 2, \cdots, S\}$) based on such features f_d ($f_d = E_d(x)$). It indicates which source domain the input image comes from. We define the domain classification loss L_{dom} as follows.

$$L_{dom} = -\sum_{j=1}^{S}\sum_{i=1}^{N_j} m_j * \log(C_d(E_d(x_i^j))), \tag{1}$$

where m_j denotes the ground truth of the image domain label.

Moreover, we provide additional supervision to ensure that the domain representation learned by E_d and C_d do not contain expression information. Specifically, with the deployment of the expression classifier C_e, we let C_e take the features f_d. And the C_e is not expected to perform expression recognition on the domain features. Therefore, we provide the following auxiliary confusion loss L_{dom}^{conf}.

$$L_{dom}^{conf} = \sum_{j=1}^{S}\sum_{i=1}^{N_j} ||C_e(E_d(x_i^j)) - \frac{1}{K}||_2^2. \tag{2}$$

With the above L_{dom} and L_{dom}^{conf}, our proposed framework can disentangle the domain features from the input expression images. In addition, the deployment of E_d and C_d also facilitates expression and identity learning.

3.3 Learning Expression and Identity Representations

Since identity information is useless for cross-domain FER, we propose to decouple it, which improves the discrimination of expression features. Extracting expression and identity representations separately from the input image is the main component of our approach. With the identity and domain representations correctly disentangled from the input image, our learned expression features can be effectively applied to recognize facial expressions in unseen target domains.

For the learning of expression features, as shown in Figure 2, the expression encoder E_e and classifier C_e are deployed in our framework to achieve this goal. E_e is expected to extract expression features ($f_e = E_e(x)$). C_e classifies expressions of the input image based on the features f_e. To better separate each expression, we utilize the simplified LMCL function

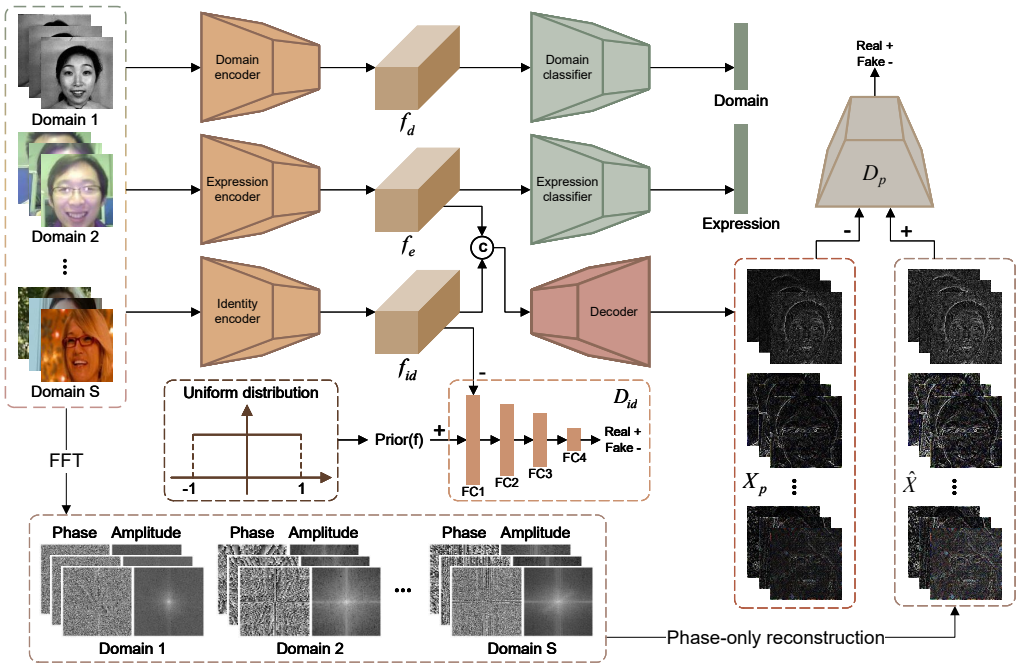


Figure 2: The framework of our method. Our network aims to extract domain features f_d, identity features f_{id}, and expression features f_e from the input facial images. To ensure the disentanglement of these features while improving the discriminative of expression features, domain encoder E_d, domain classifier C_d, expression encoder E_e, expression classifier C_e, identity encoder E_{id}, identity discriminator D_{id}, decoder De, and phase information discriminator D_p are jointly deployed. Once the training is complete, we can apply the E_e and C_e for cross-domain FER.

[Wang *et al.*, 2018] as the objective function, i.e., the expression classification loss L_{exp} is defined as follows.

$$L_{exp} = -\sum_{j=1}^{S}\sum_{i=1}^{N_j} \log\left(\frac{e^{\alpha(W_{y_i^j}^T E_e(x_i^j)-m)}}{e^{\alpha(W_{y_i^j}^T E_e(x_i^j)-m)} + \sum_{k\neq y_i^j} e^{\alpha(W_k^T E_e(x_i^j))}}\right), \tag{3}$$

where α is the hyperparameter. $W = \{W_k \mid k = 0, 1, \cdots, K\}$ represents the parameters of the C_e. The separation between the features of each expression can be further enforced by introducing the margin m.

Similarly, to further ensure that the expression features would not contain any image domain information, we provide additional supervision — the expression confusion loss L_{exp}^{conf}, defined as follows:

$$L_{exp}^{conf} = \sum_{j=1}^{S}\sum_{i=1}^{N_j} ||C_d(E_e(x_i^j)) - \frac{1}{S}||_2^2. \tag{4}$$

For the learning of identity, the identity encoder E_{id} is designed to learn the mapping of facial image to identity representation ($f_{id} = E_{id}(x)$). The high-level semantics of the image needs to be preserved after disentangling the domain representation, which facilitates the disentanglement of identity and expression information. We leverage the semantic preservation property of the Fourier phase component to provide support for preserving the high-level semantics. Note that using only Fourier phase information for cross-domain FER is not satisfactory because it still suffers from the interference of identity information. As shown in Figure 3, the facial images from different domains are processed in the following two ways. One way is the reconstruction of image with amplitude information by setting the Fourier phase component to a constant. The other is the reconstruction of image with phase information by setting the Fourier amplitude component to a constant. It can be seen that the Fourier phase information reconstructed image retains the semantics of the original image.

For example, the Fourier transformation $F(x)$ of a single channel image x is formulated as:

$$F(x)(u,v) = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} x(h,w)e^{-j2\pi\left(\frac{h}{H}u+\frac{w}{W}v\right)}. \tag{5}$$

The amplitude and phase components are denoted as:

$$A(x)(u,v) = \left[R^2(x)(u,v) + I^2(x)(u,v)\right]^{1/2}, \tag{6}$$

$$P(x)(u,v) = \arctan\left[\frac{I(x)(u,v)}{R(x)(u,v)}\right], \tag{7}$$

where $R(x)$ and $I(x)$ denote the real and imaginary part of $F(x)$, respectively. Set amplitude to be a constant. The constant we use here is the standard deviation of amplitude.

$$\hat{A}(x_i^j) = \mathrm{std}(A(x_i^j)) = \left[E(A(x_i^j) - E(A(x_i^j)))^2\right]^{1/2}. \tag{8}$$

The constant amplitude spectrum is combined with the original phase spectrum to form a new Fourier representation, as shown in the following formula:

$$F(\hat{x}_i^j)(u,v) = \hat{A}(x_i^j)(u,v) * e^{-j*P(x_i^j)(u,v)}. \quad (9)$$

It is then sent to the $F^{-1}(x)$ to generate the image, i.e., the reconstructed image with Fourier phase information is represented as follows. $F^{-1}(x)$ is the inverse Fourier transformation that maps the spectral signal back to the image space. $F(x)$ and $F^{-1}(x)$ can be calculated with the FFT algorithm [Nussbaumer, 1981] efficiently.

$$\hat{x}_i^j = F^{-1}(F(\hat{x}_i^j)(u,v)). \quad (10)$$

Then, we adopt the idea of generative adversarial and take the identity features f_{id} and expression features f_e as the input to the decoder De which is designed to recover (generate) the images reconstructed with Fourier phase information. The discriminator D_p is designed to distinguish the fake images generated by De from the real Fourier phase information reconstructed images. Besides, De and D_p assist the expression encoder E_e and the identity encoder E_{id} in learning the disentanglement representation from facial images, i.e., their deployment contributes to the disentanglement of f_{id} and f_e. The min-max objective function is defined as follows:

$$\min_{De}\max_{D_p} E[\log D_p(\hat{x}_i^j)] + E[\log(1 - D_p(De(E_e(x_i^j), E_{id}(x_i^j))))]. \quad (11)$$

Furthermore, we provide phase information similarity loss to ensure that the output image of decoder De shares expression and identity representations with the input phase information reconstructed image.

$$L_p^{sim} = \sum_{j=1}^{S}\sum_{i=1}^{N_j} ||\hat{x}_i^j - De(E_e(x_i^j), E_{id}(x_i^j))||_2^2. \quad (12)$$

The identity discriminator D_{id} imposes a uniform distribution on the identity features f_{id}, which contributes to smooth the expression transformation. Assume that $Prior(f)$ is the prior distribution and $f_{id}^* \sim Prior(f)$ represents the random sampling process from $Prior(f)$. The min-max objective function adopted is presented in the following formula:

$$\min_{E_{id}}\max_{D_{id}} E[\log D_{id}(f_{id}^*)] + E[\log(1 - D_{id}(E_{id}(x_i^j)))]. \quad (13)$$

Similar to the design of domain and expression features, we need to ensure that the identity representation learned by the E_{id} does not contain domain and expression information. With the deployment of C_d and C_e, we propose the following identity confusion loss L_{id}^{conf}:

$$L_{id}^{conf} = \sum_{j=1}^{S}\sum_{i=1}^{N_j}(||C_e(E_{id}(x_i^j)) - \frac{1}{K}||_2^2 + ||C_d(E_{id}(x_i^j)) - \frac{1}{S}||_2^2). \quad (14)$$

Together with the above classification loss, adversarial loss, similarity loss, and confusion loss, we train our proposed framework to be able to disentangle identity and expression representations. Finally, the expression representation we learned is domain invariant and discriminative owing to the disentanglement of the image domain and identity information. Therefore, our model can be effectively generalized to FER across different domains.

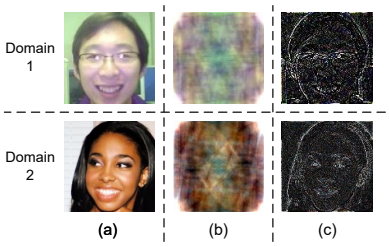


Figure 3: (a) Original images from different domains. (b) Reconstruction of images with Fourier amplitude information (set the phase component to a constant). (c) Reconstruction of images with Fourier phase information (set the amplitude component to a constant).

In summary, our overall objective function is defined as follows:

$$\begin{aligned}\min_{E_{id},De}\max_{D_{id},D_p} & E[\log D_{id}(f_{id}^*)] + E[\log(1 - D_{id}(E_{id}(x_i^j)))] \\ & + E[\log D_p(\hat{x}_i^j)] + E[\log(1 - D_p(De(E_e(x_i^j), E_{id}(x_i^j))))] \\ & + L_{exp} + L_{dom} + L_{exp}^{conf} + L_{dom}^{conf} + L_{id}^{conf} \\ & + \lambda L_p^{sim} + \tau TV(De(E_e(x_i^j), E_{id}(x_i^j))),\end{aligned} \quad (15)$$

where $TV(\cdot)$ represents the total variation that effectively eliminates ghost artifacts. The coefficients λ and τ balance smoothness and resolution. The network is trained and updated by the above formula. Once the training of our network architecture is complete, only the expression encoder E_e and classifier C_e will be utilized to perform cross-domain FER. That is, E_e is applied to extract domain-invariant and discriminative expression features. Then those features are fed into C_e for facial expression prediction.

4 Experiments

4.1 Datasets

JAFFE [Lyons *et al.*, 1998]: The JAFFE (denoted as J) is a laboratory-controlled facial expression database that contains 213 samples from 10 Japanese females. Each person has 3-4 images that are annotated with seven basic facial expressions (anger/AN, disgust/DI, fear/FE, happiness/HA, neutral/NE, sadness/SA and surprise/SU). We follow previous works [Chen *et al.*, 2021] to use the entire dataset for the training and test sets.

Oulu-CASIA [Zhao *et al.*, 2011]: The laboratory-controlled Oulu-CASIA (denoted as O) is a database of 2,880 image sequences from 80 subjects captured with two imaging systems under three different illumination conditions. These subjects are labeled with seven basic facial expressions. As in [Li *et al.*, 2022], we use 1,920 images from Oulu-CASIA VIS which is obtained by selecting the last three frames and the first frame (neutral expression) from 480 videos with the VIS system under normal indoor illumination.

RAF-DB [Li *et al.*, 2017b]: The RAF-DB (denoted as R) is a real-world database consisting of around 30,000 facial

images annotated with seven basic or eleven compound expressions. In the experiment, we select images labeled with seven basic expressions, 12,271 of which are used for training and 3,068 for testing. In addition, RAF-DB 2.0 [Li and Deng, 2020] (denoted as R2.0) is an extension of the current RAF-DB database. A subset of R2.0 contains 14, 216 images that are utilized as source data in the experimental comparisons.

SFEW 2.0 [Dhall *et al.*, 2011]: The in-the-wild database SFEW 2.0 (denoted as S) is the benchmark dataset for the SReco sub-challenge in EmotiW 2015 [Dhall *et al.*, 2015]. Each image is assigned one of the seven basic expressions. We use 958 images for training and 436 images for testing.

4.2 Experimental Results

Leave-One-Domain-Out Results

Following [Li *et al.*, 2017a], we first perform the leave-one-domain-out evaluation. Specifically, we train our model using labeled images from multiple source domains and choose the best model on the validation splits of the training set. For testing, we evaluate the selected model on a retained target domain. For example, "O&R&S to J" means that O, R, and S are treated as the source domains, and J is used as the target domain. As shown in Table 1, we provide the experimental results for different backbones (ResNet-18 and ResNet-50). We use a network consisting of the E_e and C_e as the baseline. And we compare our approach with some state-of-the-art DG methods that are not specifically designed for cross-domain FER. The superiority of our approach demonstrates that training model to disentangle expression and identity features can improve its performance on unseen domain images.

Method	O&R&S to J	J&R&S to O	J&O&S to R	J&O&R to S
ResNet-18				
Baseline	46.54	50.01	56.05	40.04
Jigen [Carlucci *et al.*, 2019]	51.77	55.03	58.99	41.93
DDAIG [Zhou *et al.*, 2020]	54.69	56.75	59.87	42.40
CSD [Piratla *et al.*, 2020]	51.09	55.85	59.13	42.27
FACT [Xu *et al.*, 2021]	57.32	58.41	59.99	43.38
CIRL [Lv *et al.*, 2022]	59.07	59.96	60.34	45.61
Ours	70.06	62.95	70.95	57.68
ResNet-50				
Baseline	49.87	52.59	58.01	41.96
Jigen [Carlucci *et al.*, 2019]	52.04	55.97	59.93	42.37
DDAIG [Zhou *et al.*, 2020]	55.11	57.62	60.31	43.41
CSD [Piratla *et al.*, 2020]	52.01	56.74	60.08	43.10
FACT [Xu *et al.*, 2021]	58.63	59.77	61.18	45.13
CIRL [Lv *et al.*, 2022]	60.03	60.75	62.96	47.32
Ours	**71.72**	**67.83**	**74.94**	**60.12**

Table 1: Comparison with the state-of-the-art DG methods on the different target domains (%).

Comparison with Other Cross-Domain FER Methods

Following the setup of previous works [Chen *et al.*, 2021; Li *et al.*, 2022], we compare the proposed method with other cross-domain FER approaches, and the results are shown in Table 2. Since our model uses domain labels, the training set requires multi-source domain data. To be consistent with the size of training set in other methods, we select a subset of the multi-source domain (R and O) to be our training set. Also, the selected source and target domains are not overlapped. It can be seen that our method outperforms the existing methods. Other cross-domain FER methods still have limited cross-domain recognition performance due to their reliance on identity information. Our method achieves impressive results even though the target domain is not accessed during training. The comparison shows that our proposed model can learn expression features with strong discriminative ability by disentangling expression-irrelevant information. Therefore, the cross-domain FER performance is improved.

Method	Backbone	Source set	J	S
DFA [Zhu *et al.*, 2016]	ResNet-18	R	42.25	38.30
	ResNet-50	R	44.44	43.07
LPL [Li *et al.*, 2017b]	ResNet-18	R	53.99	49.31
	ResNet-50	R	53.05	48.85
FTDNN [Zavarez *et al.*, 2017]	ResNet-18	R	50.23	49.31
	ResNet-50	R	52.11	47.48
DETN [Li and Deng, 2018]	ResNet-18	R	52.11	42.25
	ResNet-50	R	55.89	49.40
ICID [Ji *et al.*, 2019]	ResNet-18	R	48.83	47.02
	ResNet-50	R	50.70	48.85
AGRA [Chen *et al.*, 2021]	ResNet-18	R	61.03	52.75
	ResNet-50	R	61.50	56.43
		R2.0	62.44	-
ECAN [Li and Deng, 2020]	ResNet-18	R	52.11	48.21
	ResNet-50	R	57.28	52.29
DMSRL-RF [Li *et al.*, 2022]	ResNet-50	R2.0	68.54	-
Ours	ResNet-18	Subset of (R + O)	64.57	53.84
	ResNet-50	Subset of (R + O)	**70.03**	**57.55**

Table 2: Comparison with other cross-domain FER methods on the different target domains (%).

Ablation Study

To demonstrate the importance of each module in our proposed framework, we conduct the following ablation studies on different target domains. As shown in Table 3, we provide experimental results of models "a" (Baseline (E_e, C_e)), "b" (Baseline+E_d, C_d), "c" (Baseline+E_{id}, D_{id}, D_p, De), and "f" (Ours), respectively (the backbone is ResNet-50). The comparison of models "b" and "a" confirms that the disentanglement of image domain features assists our model in extracting domain-invariant facial content. The comparison of models "c" and "a" verifies that the disentanglement of identity information helps our model focus on facial expression, thus improving the discriminative ability of expression

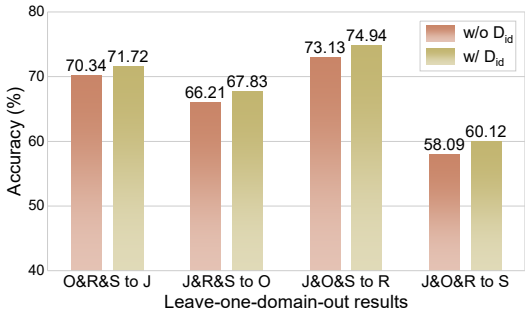


Figure 4: The influence of our introduced D_{id}. And the importance of the D_{id} for our model can be observed.

Method	O&R&S to J	J&R&S to O	J&O&S to R	J&O&R to S
Baseline (E_e, C_e)	49.87	52.59	58.01	41.96
Baseline+E_d, C_d	65.48	60.03	67.56	53.95
Baseline+ E_{id}, D_{id}, D_p, De	67.75	63.22	70.63	56.54
w/o confusion loss	69.27	65.04	71.93	57.81
w/o $L_p^{sim} + TV(\cdot)$	71.69	67.68	74.75	59.98
Ours	**71.72**	**67.83**	**74.94**	**60.12**

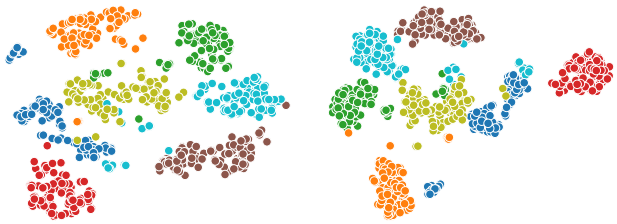
Table 3: Analysis of the modules (and losses) for our network architecture. The evaluation reveals that each module (and loss) contributes to the effectiveness of our overall architecture (%).

features. Moreover, the effectiveness of leveraging Fourier phase information to help disentangle identity features is also demonstrated. The comparison of models "f" and "a" shows that our proposed representation disentanglement network can learn domain-invariant and discriminative expression features, thus bringing significant improvements. Models "d" (w/o confusion loss) and "e" (w/o $L_p^{sim}+TV(\cdot)$) are ablation studies of the loss function, illustrating the effectiveness of our introduction of confusion loss, similarity loss, and $TV(\cdot)$. In addition, the identity discriminator D_{id} is introduced to smooth the transformation of expressions. As shown in Figure 4, we provide the experimental results before and after the introduction of the D_{id}. The effectiveness of introducing the D_{id} can be observed.

Visualization

We visualize the extracted expression features to reflect their domain invariance. Furthermore, to enforce the disentanglement among features, the auxiliary confusion loss is introduced. Therefore, we provide the t-SNE [Van der Maaten and Hinton, 2008] visualization results as in Figure 5. Here we choose R as the target domain and perform leave-one-domain-out experiments (J&O&S to R, backbone is ResNet-50). Specifically, we train without and with confusion loss, respectively, and then obtain the trained E_e and C_e. During testing, the images in the target domain (we choose around 100 facial images per expression) is randomly selected as the input of the trained E_e to extract expression features. The acquired features are visualized separately in Figure 5. We can see that our approach can well capture the domain-invariant expression features. Also, the introduction of the confusion loss helps to improve the performance of cross-domain FER.

Our method extracts expression features by disentangling identity and domain information. To further verify its effectiveness, we use the Grad-CAM [Selvaraju *et al.*, 2017] algorithm to obtain the class activation mapping visualizations. We select O and R as the target domain for the leave-one-domain-out experiments (J&R&S to O and J&O&S to R, backbone is ResNet-50). The class activation maps for each expression in different domains is shown in Figure 6. The image regions that the model focuses on during the cross-domain FER can be seen. The (a) and (b) denote models "b" and "f" (Ours) in Table 3. We find that model "b" can focus on the facial region without the influence of domain information. And model "f" can locate more expression-related potential interesting areas by disentangling the identity features. That



(a) w/o confusion loss (b) w/ confusion loss

Figure 5: The t-SNE visualization of expression features under the cases without and with confusion loss. Different colors represent different expressions. It is clear that our method has good cross-domain FER performance. And confusion loss can contribute to improving the discrimination of expression features.

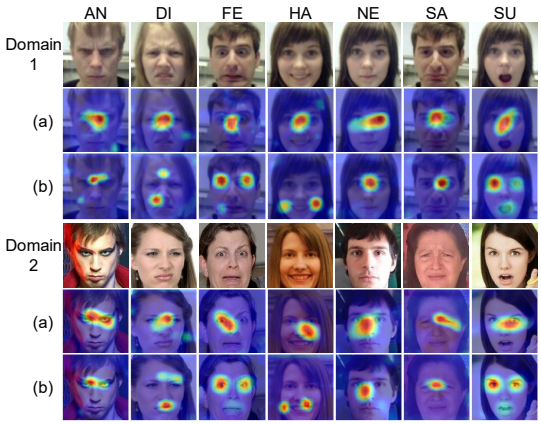


Figure 6: Grad-CAM visualization of different expressions on different domains. Each column displays each facial expression. Domains 1 and 2 (rows 1 and 4) denote the original facial images (after face detection) from different domains (O and R). The (a) (rows 2 and 5) and (b) (rows 3 and 6) show the test results for models "b" and "f" (Ours) in Table 3. Our method disentangles the domain and identity information so that the learned expression features exhibit good domain invariance and discrimination.

is, our whole disentanglement model can focus on domain-invariant and more discriminative facial regions. Therefore, the performance of cross-domain FER is improved.

5 Conclusion

In this paper, we attempt to solve the more challenging task of cross-domain FER not being able to access the target domain data. We propose a representation disentanglement model capable of extracting domain, identity, and expression features based on the idea of DG. To ensure the disentanglement among features, we use the semantic preservation property of the Fourier phase component to provide support for preserving the high-level semantics of the image. In addition, we introduce classification loss, adversarial loss, similarity loss, and confusion loss, respectively. And the training is implemented by our designed objective loss function. Extensive experiments on multiple datasets demonstrate the effectiveness of our method.

Acknowledgements

This work is partially supported by National Natural Science Foundation of China (No.62206116), Key Research and Development Program of Hubei Province (2022BAA079), Major Science and Technology Projects of Jilin Province (20210301030GX), and Leading innovation and entrepreneurship team of Zhejiang (2019R01015).

References

[Carlucci *et al.*, 2019] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *CVPR*, pages 2229–2238, 2019.

[Chattopadhyay *et al.*, 2020] Prithvijit Chattopadhyay, Yogesh Balaji, and Judy Hoffman. Learning to balance specificity and invariance for in and out of domain generalization. In *ECCV*, pages 301–318, 2020.

[Chen *et al.*, 2016] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *NIPS*, 2016.

[Chen *et al.*, 2021] Tianshui Chen, Tao Pu, Hefeng Wu, Yuan Xie, Lingbo Liu, and Liang Lin. Cross-domain facial expression recognition: A unified evaluation benchmark and adversarial graph learning. *TPAMI*, pages 9887–9903, 2021.

[Dhall *et al.*, 2011] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *ICCV workshops*, pages 2106–2112, 2011.

[Dhall *et al.*, 2015] Abhinav Dhall, OV Ramana Murthy, Roland Goecke, Jyoti Joshi, and Tom Gedeon. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 423–426, 2015.

[Gan *et al.*, 2016] Chuang Gan, Tianbao Yang, and Boqing Gong. Learning attributes equals multi-source domain generalization. In *CVPR*, pages 87–97, 2016.

[Hansen and Hess, 2007] Bruce C Hansen and Robert F Hess. Structural sparseness and spatial phase alignment in natural scenes. *JOSA A*, pages 1873–1885, 2007.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Huang *et al.*, 2021a] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Fsdr: Frequency space domain randomization for domain generalization. In *CVPR*, pages 6891–6902, 2021.

[Huang *et al.*, 2021b] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Rda: Robust domain adaptation via fourier adversarial attacking. In *ICCV*, pages 8988–8999, 2021.

[Ji *et al.*, 2019] Yanli Ji, Yuhan Hu, Yang Yang, Fumin Shen, and Heng Tao Shen. Cross-domain facial expression recognition via an intra-category common feature and inter-category distinction feature fusion network. *Neurocomputing*, pages 231–239, 2019.

[Kang *et al.*, 2019] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *CVPR*, pages 4893–4902, 2019.

[Li and Deng, 2018] Shan Li and Weihong Deng. Deep emotion transfer network for cross-database facial expression recognition. In *2018 24th International Conference on Pattern Recognition*, pages 3092–3099, 2018.

[Li and Deng, 2020] Shan Li and Weihong Deng. A deeper look at facial expression dataset bias. *IEEE Transactions on Affective Computing*, pages 881–893, 2020.

[Li *et al.*, 2017a] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, pages 5542–5550, 2017.

[Li *et al.*, 2017b] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *CVPR*, pages 2852–2861, 2017.

[Li *et al.*, 2018] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *CVPR*, pages 5400–5409, 2018.

[Li *et al.*, 2021] Yingjian Li, Yingnan Gao, Bingzhi Chen, Zheng Zhang, Lei Zhu, and Guangming Lu. Jdman: Joint discriminative and mutual adaptation networks for cross-domain facial expression recognition. In *ACM MM*, pages 3312–3320, 2021.

[Li *et al.*, 2022] Yingjian Li, Zheng Zhang, Bingzhi Chen, Guangming Lu, and David Zhang. Deep margin-sensitive representation learning for cross-domain facial expression recognition. *IEEE Transactions on Multimedia*, pages 1359–1373, 2022.

[Lv *et al.*, 2022] Fangrui Lv, Jian Liang, Shuang Li, Bin Zang, Chi Harold Liu, Ziteng Wang, and Di Liu. Causality inspired representation learning for domain generalization. In *CVPR*, pages 8046–8056, 2022.

[Lyons *et al.*, 1998] Michael Lyons, Shigeru Akamatsu, Miyuki Kamachi, and Jiro Gyoba. Coding facial expressions with gabor wavelets. In *Proceedings Third IEEE international conference on automatic face and gesture recognition*, pages 200–205, 1998.

[Muandet *et al.*, 2013] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *ICML*, pages 10–18, 2013.

[Nussbaumer, 1981] Henri J Nussbaumer. *The fast Fourier transform*. Springer Berlin Heidelberg, 1981.

[Oppenheim and Lim, 1981] Alan V Oppenheim and Jae S Lim. The importance of phase in signals. *Proceedings of the IEEE*, pages 529–541, 1981.

[Piratla *et al.*, 2020] Vihari Piratla, Praneeth Netrapalli, and Sunita Sarawagi. Efficient domain generalization via common-specific low-rank decomposition. In *ICML*, pages 7728–7738, 2020.

[Recht *et al.*, 2019] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, pages 5389–5400, 2019.

[Selvaraju *et al.*, 2017] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017.

[Shao *et al.*, 2019] Rui Shao, Xiangyuan Lan, Jiawei Li, and Pong C Yuen. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *CVPR*, pages 10023–10031, 2019.

[Tzeng *et al.*, 2017] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, pages 7167–7176, 2017.

[Van der Maaten and Hinton, 2008] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, pages 2579–2605, 2008.

[Wang *et al.*, 2018] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *CVPR*, pages 5265–5274, 2018.

[Wang *et al.*, 2022] Zhuo Wang, Zezheng Wang, Zitong Yu, Weihong Deng, Jiahong Li, Tingting Gao, and Zhongyuan Wang. Domain generalization via shuffled style assembly for face anti-spoofing. In *CVPR*, pages 4123–4133, 2022.

[Xu *et al.*, 2021] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *CVPR*, pages 14383–14392, 2021.

[Yang *et al.*, 2020] Yanchao Yang, Dong Lao, Ganesh Sundaramoorthi, and Stefano Soatto. Phase consistent ecological domain adaptation. In *CVPR*, pages 9011–9020, 2020.

[Zavarez *et al.*, 2017] Marcus Vinicius Zavarez, Rodrigo F Berriel, and Thiago Oliveira-Santos. Cross-database facial expression recognition based on fine-tuned deep convolutional network. In *2017 30th SIBGRAPI Conference on Graphics, Patterns and Images*, pages 405–412, 2017.

[Zhang *et al.*, 2021] Haifeng Zhang, Wen Su, Jun Yu, and Zengfu Wang. Weakly supervised local-global relation network for facial expression recognition. In *IJCAI*, pages 1040–1046, 2021.

[Zhang *et al.*, 2022] Hanlin Zhang, Yi-Fan Zhang, Weiyang Liu, Adrian Weller, Bernhard Schölkopf, and Eric P Xing. Towards principled disentanglement for domain generalization. In *CVPR*, pages 8024–8034, 2022.

[Zhao *et al.*, 2011] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z Li, and Matti PietikäInen. Facial expression recognition from near-infrared videos. *Image and vision computing*, pages 607–619, 2011.

[Zhou *et al.*, 2020] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *AAAI*, pages 13025–13032, 2020.

[Zhou *et al.*, 2021a] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *TPAMI*, pages 4396–4415, 2021.

[Zhou *et al.*, 2021b] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*, 2021.

[Zhu *et al.*, 2016] Ronghang Zhu, Gaoli Sang, and Qijun Zhao. Discriminative feature adaptation for cross-domain facial expression recognition. In *2016 International Conference on Biometrics*, pages 1–7, 2016.