2024 18th International Conference on Automatic Face and Gesture Recognition (FG)

Expression-aware Masking and Progressive Decoupling for Cross-database Facial Expression Recognition

Tao Zhong[1*], Xiaole Xian[1*], Zihan Wang[1], Weicheng Xie[1,2,3†] and Linlin Shen[1,2,3]
[1] Computer Vision Institute, School of Computer Science & Software Engineering, Shenzhen University
[2] Shenzhen Institute of Artificial Intelligence and Robotics for Society
[3] Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University

Abstract— **Cross-database facial expression recognition (CD-FER) has been widely studied due to its promising applicability in real-life situations, while the generalization performance is the main concern in this task. For improving cross-database generalization, current works frequently resort to masked autoencoder (MAE) to learn the expression representation in an unsupervised manner, and disentanglement of expression and domain features. (i) For MAE, current algorithms mainly employ random masking, and leverage the reconstruction of these masked regions to enable networks to learn the expression representation. However, these masked regions are expression-irrelevant, can not well reflect the characteristics of expression, thus are not efficient enough in representation learning. To this end, we propose an expression-aware masking in MAE to improve the learning efficiency of expression representation, by guiding MAE to mask out expression-aware regions during training. (ii) For disentanglement of expression and domain features, current algorithms realize it mainly in the deep layers. However, the coupling of these features in the shallow layers are rarely concerned, which may largely affect the disentanglement performance in deep layers. Thus, we propose a progressive decoupler to disentangle these features block by block, to use the feature disentanglement in shallow layers to facilitate that in deep layers. Extensive quantitative and qualitative results on multiple expression datasets show that our method can largely outperform the state of the arts in terms of cross-database generalization performance.**

I. Introduction

Facial expression recognition (FER) has received widespread attention in recent years due to its important role in human-computer interaction. However, the performance of FER in the wild scenarios is sensitive to the factors of races, genders, illumination, head pose changes, occlusions, as well as the domain gaps. This domain gap problem can be mitigated by some existing domain adaptation (DA) methods, while they require unlabeled testing data during the training phase. By contrast, domain generalization (DG) methods received increasing attention since the target domain dataset is not necessary during training.

However, the limited labeled expression data and its fixed collection environment pose a bottleneck for expression representation learning. Masked autoencoder (MAE) [11] adopted an unsupervised paradigm and used an encoder-decoder architecture for masked feature modeling. It randomly masks out a ratio, e.g. 75% of the patches in the image and trains an autoencoder to learn how to effectively reconstruct the masked patches by exploring the correlations between the unmasked and masked regions of the image. This approach achieves competitive performances in many fields [6], [12], [13], for example, MAE-DFER [36] and Audio-MAE [13] use MAE to explore unlabeled video and audio datasets, respectively, in a self-supervised manner.

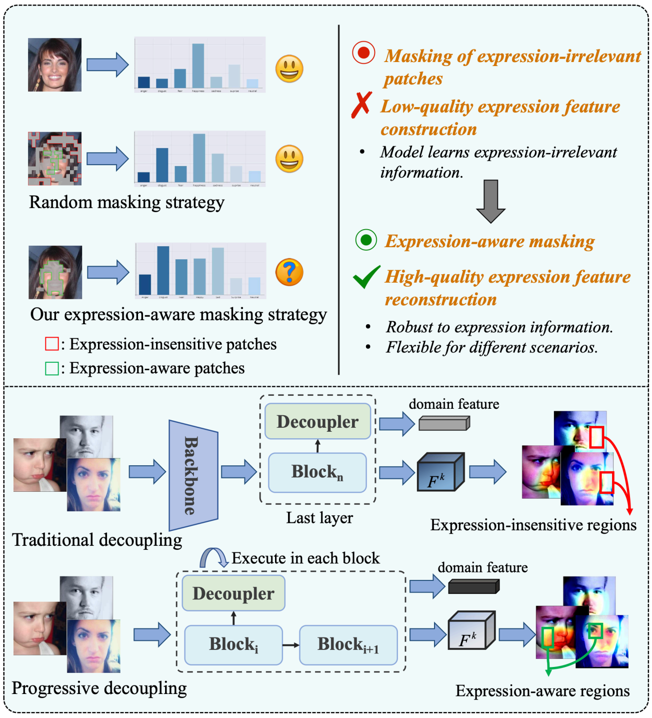
Random masking strategy
Our expression-aware masking strategy
□: Expression-insensitive patches
□: Expression-aware patches
Masking of expression-irrelevant patches
✗ Low-quality expression feature construction
Model learns expression-irrelevant information.
Expression-aware masking
✓ High-quality expression feature reconstruction
Robust to expression information.
Flexible for different scenarios.
Backbone
Decoupler
Block$_n$
Last layer
domain feature
F^k
Expression-insensitive regions
Traditional decoupling
Execute in each block
Decoupler
Block$_i$
Block$_{i+1}$
domain feature
F^k
Progressive decoupling
Expression-aware regions

Fig. 1. The motivation of our Expression semantic guidance masking (ESGM) and Progressive decoupler of expression and domain (PDED).

But the above works are all random masking, our insight is that the reconstruction of MAE is to serve our downstream cross-domain facial expression recognition, and the information density is concentrated at the facial regions relevant to expressions. As shown in the top of Fig. 1, random masking could potentially result in masking many facial low relevant patches while retaining expression-related patches, so there

The work was supported by Natural Science Foundation of China under grants no. 62276170, 82261138629, the Science and Technology Project of Guangdong Province under grants no. 2023A1515011549, 2023A1515010688, the Science and Technology Innovation Commission of Shenzhen under grant no. JCYJ20220531101412030.

[*] Equal contribution.

[†] Corresponding author.

is no guarantee that the most informative patches will be properly focused on. To this end, we propose expression semantic guidance masking (ESGM), a masking strategy to ensure that most expression-related patches can be masked.

In addition to the limitation of the labeled data in expression representation learning, CDFER also suffers from the entanglement of expression and domain features. For extracting domain-invariant expression features, existing decoupling methods are mainly divided into three categories, i.e., single decoupling [5], cyclic decoupling [50], and parallel decoupling [35]. Single decoupling only decoupled once, which may result in incomplete decoupling. Cyclic decoupling decouples the same feature multiple times to enhance the effect of separating domain and expression features. Parallel decoupling uses multiple decouplers to decouple the same feature, but the obtained features may be redundant. Previous works only performed decoupling on deep features of a network, as shown in the bottom of Fig. 1, they produce inappropriate attention in expressions. For the reasons, they may ignore the entanglement of expression and domain features in shallow layers, which affect the decoupling capacity in deep layers. Thereby, we proposed progressive decoupler of expression and domain (PDED) to disentangle these features in both shallow and deep layers, so as to use their disentanglement in shallow layers to facilitate learning domain-invariant expression features in deep layer.

To address the limitations discussed above, we propose a novel two-stage framework for cross-domain facial expression recognition. In the pre-training stage, we employ an encoder-decoder architecture for self-supervised pre-training. A novel masking strategy, i.e. expression semantic guidance masking (ESGM) is designed to enable the model to selectively attend to expression-aware regions during expression information masking and reconstruction, so as to specifically learn the expression feature representation. In this way, we can distill expression-specific patterns from different unlabeled facial datasets, avoiding over-reliance on the mere source training dataset.

In addition to the representation of expression sensitive cues via expression-aware masking and reconstruction, we take into account the impact of shallow-layer feature entanglement on that on deep layer, and propose a progressive decoupler to separate domain and expression features block by block. In this way, the network can obtain domain-invariant expression features at deep blocks via progressive disentanglement from shallow layers that are more conducive to CDFER.

The contributions of this paper are summarized as follows

- We propose a novel masking strategy to enable an encoder-decoder to focus more on facial expression information for masked modeling, and eventually obtain an encoder that is aware of expression cues.
- Taking into account the coupling of expression and domain features in shallow layers will impair their disentanglement in deep layers, we propose a progressive decoupling module that disentangles domain-invariant expression features and domain features block by block.
- Our method outperforms the related state of the arts for the task of cross-database FER in terms of domain generalization performance.

II. Related Works

A. Masked Autoencoder

Masked modeling [11] learns robust feature representations through masking and reconstruction, and has achieved great success in natural language processing (like BERT [7]) and computer vision. Masked AutoEncoder introduced an asymmetric encoder-decoder architecture for masked image modeling, which was inspired by the denoising autoencoder [39] and image inpainting [32], and was used in the field of computer vision after the success of BERT, such as image recognition [19], video recognition [38], and point cloud [30]. BEiT [2] is one of the pioneering works to apply image masking and reconstruction for downstream tasks. SemMAE [19] limits the mask area through attention map. VideoMAE [38] masks some video clips in the time dimension. These methods all use random masking, which do not differentiate the patches according to the characteristics of the considered task. Latent-OFER [18] used anomaly classification to select the occluded mask, but lacking a loss to constrain it.

In this work, we propose the expression semantic guidance masking (ESGM) strategy to replace random masking in MAE for reconstructing expression-related regions, so as to obtain an encoder robust to expression representation via a self-supervised training.

B. Cross-database Facial Expression Recognition (CDFER)

Two categories of algorithms are frequently considered in cross-database expression recognition, i.e., domain adaptation [22], [25], [41] and domain generalization [14], [29], [40], [44]. Li et al. [22] propose ECAN to learn domain-invariant and discriminative feature representations. Xie et al. [41] propose the AGRA framework, which combines graph representation propagation with adversarial learning. However, these domain adaptation methods require unlabeled data of the target domain.

By contrast, domain generalization does not need target domain data during training. TDTLN [43] devised the cross-database-specific discriminative features, and Ji et al. proposed ICID [14] to learn both intra-category common features and inter-category discriminative features. Zou et al. [50] proposed Learn-to-Decompose (LD) modules to reconstruct the transferable expression feature. However, above works only perform the expression and domain decoupling in deep layers without considering the entanglement in shallow layers, which may harm the decoupling capacity in deep layers. In this work, we propose to decouple expression and domain features at different network layers via our proposed progressive disentanglement of these features.

III. Methodology

In this section, we introduce the main modules of our proposed framework in Fig. 2, including Expression semantic guidance masking (ESGM) and Progressive decoupler of

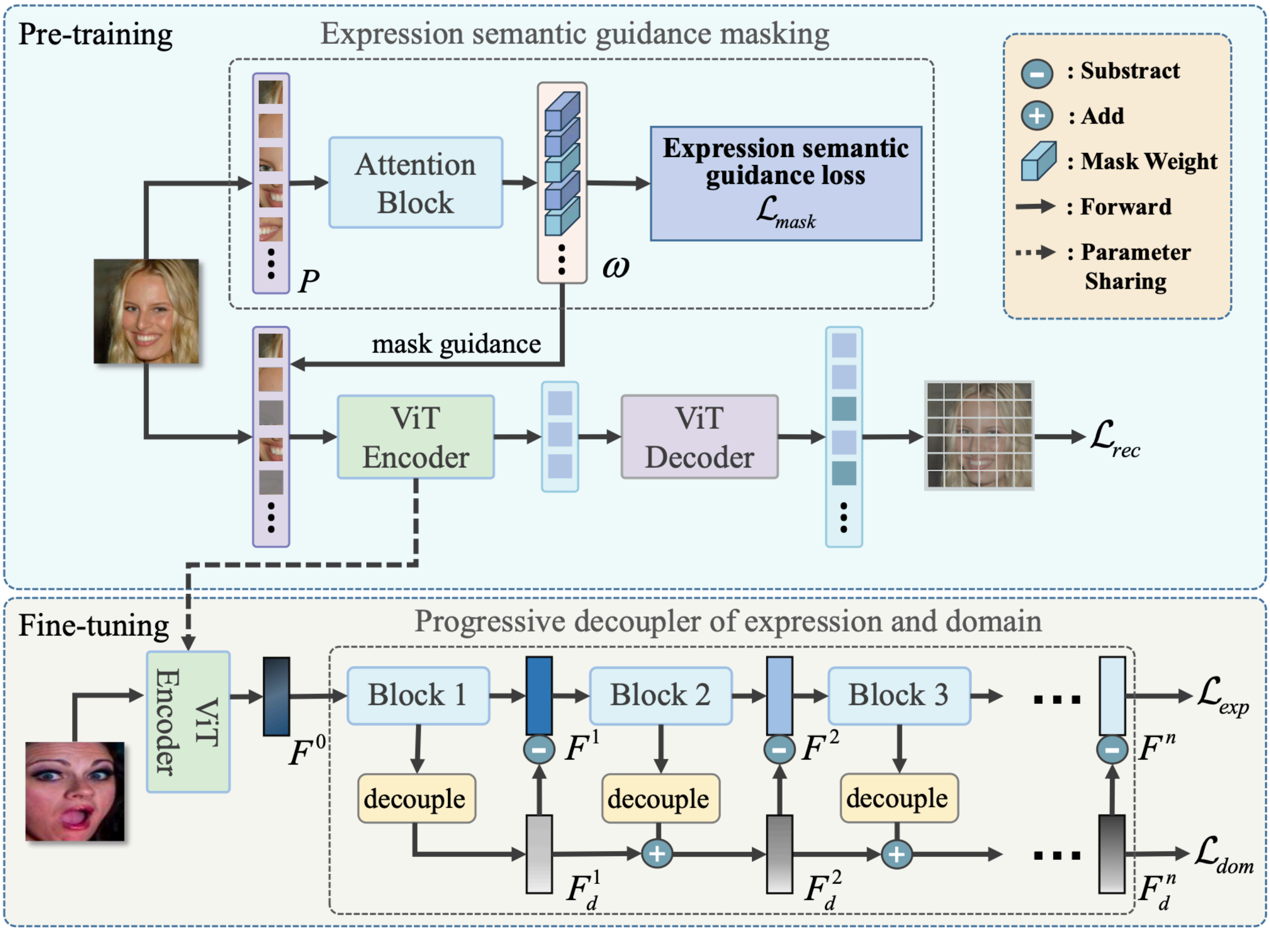

Fig. 2. Our method consists of the modules of Expression semantic guidance masking (ESGM), Progressive Decoupler of expression and domain (PDED).

expression and domain (PDED), corresponding to the pre-training and fine-tuning stages, respectively. ESGM aims to train an encoder that has learned the facial expression-related information from the unlabeled facial image with the assistance of attention block adaptively in pre-training stage. In the fine-tuning stage, PDED further decouples the expression and domain features block by block, and uses domain-invariant expression features for classification.

A. Expression semantic guidance masking (ESGM)

As our motivation stated, random masking does not take into account the characteristics of face images. Therefore, we design a novel adaptive masking strategy that allows us to guide model pre-training under the supervision of face parsing segmentation, as shown in Fig. 3. Specifically, following ViT [9], we divide a face image input x_f into a set of several non-overlapping patches and map them into tokens P as:

$$P = \left[P^{(0)}, P^{(1)}, \cdots, P^{(N_p-1)}\right], \tag{1}$$

where $N_p = (H \times W)/(ps \times ps)$ is the number of patches, H, W and ps are the height, width and patch size, respectively, and $P \in \mathbb{R}^{N_p \times d}$. We then use a Transformer encoder [9] which consists of Multi-head Self-Attention (MSA) and

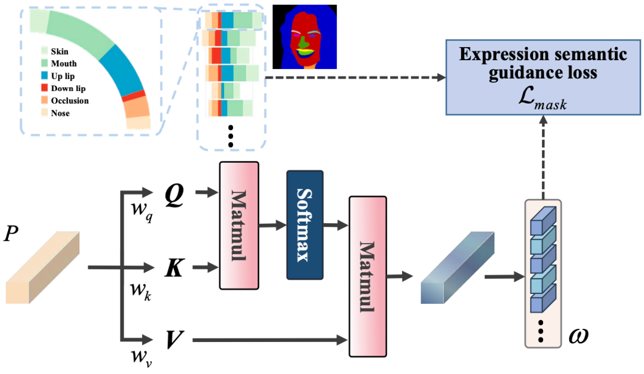

Fig. 3. Mechanism of the expression semantic guidance masking (ESGM). "Matmul" operation denotes the matrix multiplication.

Multi-Layer Perception (MLP) with skip connections to get the mask weight ω.

Firstly, the input $P \in \mathbb{R}^{N_p \times d}$ is linearly transformed to queries Q, keys K, and values V as follows

$$[Q, K, V] = P[w_q, w_k, w_v], \tag{2}$$

where $w_q, w_k \in \mathbb{R}^{\mathrm{d} \times \mathrm{d}_k}, w_v \in \mathbb{R}^{\mathrm{d} \times \mathrm{d}_v}$ are learnable variables.

In MSA, self-attention operations are performed several times in parallel, and linearly embeds their concatenated outputs. Eventually, the weighting vector ω is obtained by inputting the features into a pooling layer and several linear layers as follows

$$\begin{gathered}\omega = MLP(\mathbb{GAP}(A \cdot V)), \\ \text{with } A = Softmax(\frac{Q \cdot K^T}{\sqrt{d_k}}),\end{gathered} \tag{3}$$

where $\cdot$ means matrix multiplication, $\mathbb{GAP}$ is the global average pooling operation, $\sqrt{d_k}$ is a scaling factor, $\omega = \left[\omega^{(0)}, \omega^{(1)}, \cdots \omega^{(N_p-1)}\right]$ represents the possibility of the patch being masked and $\omega^{(i)} \in [0,1]$.

Then, we obtain the indexes of these patches according to their magnitudes. The indexes of the masked patches, i.e., idx_{mask} can be represented as

$$\begin{aligned}idx_{mask} &= \left[idx^{(0)}, idx^{(1)}, \cdots, idx^{(\lfloor mr \times N_p \rfloor - 1)}\right], \\ \text{with } idx &= argsort(\omega),\end{aligned} \tag{4}$$

where $argsort(\omega)$ means the indexes after ω is sorted in the descending order. The hyperparameter mr denotes the mask ratio. In Eq. (4), we aim to select a proportion of dominant patches determining by the mask ratio of mr, to reflect expression-aware cues or the domain-irrelevant features.

To make the patches corresponding to our desired masking targets, we introduce face parsing segmentation $\mathbf{S}_{parsing}$ as the guidance. Specifically, we empirically select the facial attributes that are related to expressions into a set as $S_{exp} = \{1: nose,\ 2: eye,\ 3: brow,\ 4: mouth,\ 5: lip,\ 6: skin\}$, then use an indicator map $\mathcal{M}_{exp}$ to represent whether an image pixel belongs to S_{exp}, which is formulated as:

$$\mathcal{M}_{exp}(i,j) = \begin{cases} 1, & \mathbf{S}_{parsing}(i,j) \in S_{exp} \\ 0, & otherwise \end{cases}, \tag{5}$$

where $\mathbf{S}_{pasing}(i,j)$ denotes the pixel value of i-th row and j-th column in the parsing segmentation map. Then we divided $\mathcal{M}_{exp}$ into N_p patches, i.e., $\left[\mathcal{M}_{exp}^{(0)}, \mathcal{M}_{exp}^{(1)}, \ldots, \mathcal{M}_{exp}^{(N_p-1)}\right]$. To enable networks to focus on the patches containing more expression attributes, we use $\mathcal{M}_{exp}$ to guide the mask weight ω, and the loss is formulated as:

$$\mathcal{L}_{mask} = \sum_{k=0}^{N_p-1} |\omega^{(k)} - \sum_{i=0}^{ps-1} \sum_{j=0}^{ps-1} \frac{\mathcal{M}_{exp}^{(k)}(i,j)}{ps \times ps}|, \tag{6}$$

Since a larger $\mathcal{M}_{exp}^{(k)}$ implies a higher density of expression-related information, the attention block can pay attention to the expression-related patches, corresponding to large value of ω, under the constraint of $\mathcal{L}_{mask}$. Then these patches are masked, while the unmasked patches are input into an asymmetric encoder-decoder for reconstructing the image. While the encoder only encodes the unmasked tokens, the decoder decodes both encoded unmasked tokens and masked tokens, where the mask tokens needed to be predicted are randomly initialized. In addition, positional embeddings are added to the unmasked and masked tokens to let the model capture the spatial context of the masked tokens in an image. We follow MAE [11] and utilize the MSE loss $\mathcal{L}_{rec}$ to constrain the reconstruction of the masked tokens as:

$$\mathcal{L}_{rec} = \sum_{i \in idx_{mask}} \|P^{(i)} - \mathcal{F}_{pt}(P^{(i)}, \theta_{\mathcal{F}_{pt}})\|_2^2, \tag{7}$$

where $\mathcal{F}_{pt}$ denotes the entire reconstruction network, $\theta_{\mathcal{F}_{pt}}$ represents its parameters. The total loss of the pre-training stage is then formulated as

$$\mathcal{L}_{pt} = \mathcal{L}_{rec} + \lambda_1 \mathcal{L}_{mask}, \tag{8}$$

where λ_1 is a hyperparameter.

For our masking strategy, a few epochs of warming up training are needed to enable the attention block to focus on the expression-aware facial patches. That is, this part of the pre-training requires using datasets of face parsing results as the supervision cues. After ω has been learned in these epochs, the trained attention block can mask the expression-aware patches without the supervision of face parsing data in the latter epochs of the pre-training. Thus, the encoder after the entire pre-training could reduce the sensitivity of the expression representation to a specific domain.

B. Progressive decoupler of expression and domain (PDED)

In addition to the ESGM module that models expression-aware features, we further disentangle expression and domain features to adapt to CDFER. However, current works mainly decouple these features in deep layers, and most of them neglect the entanglement of these features in the shallow layers. Thus, we propose progressive decoupler of expression and domain (PDED) to obtain domain-invariant expression features during the fine-tuning stage, i.e. this disentanglement is performed at each block of networks, as shown in Fig. 2.

First, we input the expression image x_e (original domain) or x_a (data augmentation e.g. Gaussian blurring for single source domain or the domains other than the source domain in multi-source domains) to the $Encoder$ obtained in the pre-training stage, and get the feature F^0. To obtain domain-invariant expression features, we then remove domain-related features at each block as:

$$F^{i+1} = F^i - F_d^i, \tag{9}$$

where F^{i+1} is the input of $(i+1)$-th block, and each domain feature F_d^i is represented with a convolution operator with the same channels before and after this convolution.

By eliminating the influence of domain features at each block, we can finally obtain disentangled expression features and domain features as follows

$$\begin{cases} F_{dom} = \sum_{i=1}^{n} F_d^i \\ F_{exp} = F^n \end{cases}, \tag{10}$$

Meanwhile, to better extract the domain-invariant expression feature, we use the expression classification loss as the supervision, formulated as follows

$$\mathcal{L}_{exp} = -\frac{1}{M_S} \sum_{i=1}^{M_S} y_e^{(i)} \log(\mathcal{F}_{ft}(F_{exp}^{(i)}, \theta_{\mathcal{F}_{ft}})), \tag{11}$$

where M_S is the number of samples from the source domain dataset, $F_{exp}^{(i)}$ and $y_e^{(i)}$ are the expression feature of i-th training sample and its ground truth label, $\mathcal{F}_{ft}$ is the employed fine-tune network and $\theta_{\mathcal{F}_{ft}}$ denotes its parameters. In addition, the domain classification loss is also employed, which is formulated as:

$$\mathcal{L}_{dom} = -\frac{1}{M_S}\sum_{i=1}^{M_S} y_d^{(i)}\log(F_{dom}^{(i)})+(1-y_d^{(i)})\log(1-F_{dom}^{(i)}), \tag{12}$$

where $F_{dom}^{(i)}$ and $y_d^{(i)}$ are the domain feature of i-th training sample and its domain label. Eventually, we train the classifier based on the loss $\mathcal{L}_{ft}$ as follows

$$\mathcal{L}_{ft} = \mathcal{L}_{exp} + \lambda_2\mathcal{L}_{dom}, \tag{13}$$

where λ_2 is a hyperparameter. For clarity, the training procedure of our algorithm is shown in Algorithm 1.

Algorithm 1 The training procedure of our method.

```
Input: Samples of unlabeled facial image $\{x_f^{(i)}\}_{i=1}^{N}$, labeled expression images and its augmentation images $\{x_e^{(i)}, x_a^{(i)}, y_e^{(i)}\}_{i=1}^{M}$.
Output: Final model parameter $\theta_{ft}$ for the prediction.
# Expression semantic guidance masking:
while not converged do
    Input $x_f^{(i)}$, get token $P$ in Eq. (1), patch weights $\omega$ in Eq. (3) for our masking strategy;
    Mask the tokens $P$ using index $idx_{mask}$ in Eq. (4);
    Reconstruct the masked patches and update the $Encoder$ with loss in Eq. (8).
end while
# Progressive decoupler of expression and domain:
while not converged do
    Input $x_e^{(i)}$ or $x_a^{(i)}$ to the $Encoder$, get feature $F^0$;
    Decouple the features in each network block in Eq. (9), get domain feature $F_{dom}$ and expression feature $F_{exp}$ in Eq. (10);
    Update the $\theta_{ft}$ using expression classification loss $\mathcal{L}_{exp}$ in Eq. (11) and domain classification loss $\mathcal{L}_{dom}$ in Eq. (12).
end while
```

IV. Experiments

A. Experimental Setup

Dataset. We use CelebA [26], FFHQ [16], CASIA-WebFace [45] and CelebAMask-HQ [17] datasets at pre-training stage and finetune at the labeled datasets of RAFDB [21] and AffectNet [31]. Ultimately, the proposed approach is evaluated on six public databases, i.e., RAFDB [21], FER2013+ [3], SFEW2.0 [8], AffectNet [31], ExpW [48] and JAFFE [28] with seven expressions, i.e., six basic expressions and neutral. The size of input images is set as 224×224.

Implementation details. For the pre-training stage, we first utilize the MAE pretrained on ImageNet-1K as the backbone, and pre-train the model for 20 epochs on CelebAMask-HQ [17], aiming to get an attention block with finely annotated face parsing segmentation. Subsequently, we utilize CelebA [26], FFHQ [16], CASIA-WebFace [45], which has no parsing segmentation for pre-training 350 epochs. At the fine-tuning stage, we conduct the training on the source dataset and evaluate the performance on other datasets for CDFER. The networks are optimized via AdamW [27] with the learning rates of 1.5×10^{-4} and 1.0×10^{-4} in the pre-training and fine-tuning stages, respectively. Weight decay is set to 0.05, momentum is set to 0.9, and used cosine scheduler at the initial training of the two stages. The hyperparameters λ_1 in Eq. (8) and λ_2 in Eq. (13) are both set to 1. All the experiments are conducted on 4 GPU of NVIDIA 3090 with 24GB memory.

B. Comparison with the state of the arts

1) Performance under single source domain: We first evaluate the classification performance of our method, perform training on the RAFDB or AffectNet, and evaluate the performance on the remaining five databases and show the results in Table I.

Compared with ICID [14] designed specifically for cross-domain FER, Table I shows that our method achieves an improvement of more than 9.85% in terms of average accuracy, and a large margin of 6.17% when AffectNet is used for the training. Compared with other FER methods, including ActiveFER [33], MLA [1], etc., that are not specifically designed for cross-domain FER, our method achieves an improvement more than 7% on the five target domains in terms of average accuracy. Compared with other state-of-the-art (SOTA) DG methods, including PDEN [20], Sequeener [37], SADA [46], etc., that are not specially designed for FER, our method consistently outperforms these SOTAs in all five sets of experiments. For example, when RAFDB is used as the source database, our method achieves an improvement of 10.81% over the SADA [46].

For the masking strategy, MARLIN [4] selects expression patches according to their importance degree, which is similar to ours. Thus, we replace our ESGM strategy with that in MARLIN [4] for the ablation study of our masking strategy. Compared to MARLIN, our method achieved improvements of 1.05% and 0.31% on the two datasets, respectively, showing the effectiveness of our ESGM module. It's worth noting that our method does not introduce any randomness in the selection of expression face regions, which thus appear more robust in real application scenarios.

2) Performance on the testing dataset: To study the performance of our algorithm on the testing dataset of the source domain, Table II presents the accuracy of our method compared with other SOTAs. It shows that that our method can achieve the best results on the test dataset of RAFDB. Specifically, our algorithm achieves improvements of 0.94% over MAE+ResNet18 and 1.66% over EAC [47].

3) Performance under multiple source domains: We further evaluate the domain generalization performance of our method under the setting of multiple source domains, and show the results in Table III. It shows that our method

TABLE I

THE PERFORMANCE OF OUR ALGORITHM IN THE SCENARIO OF SINGLE-SOURCE DOMAIN GENERALIZATION. THE BEST RESULTS ARE LABELED IN BOLD. 'AVG' DENOTES THE AVERAGE ACCURACY. † MEANS THAT THE CODES ARE REPRODUCED BY US. ‡ MEANS REPLACING THE ESGM WITH MARLIN [4]'S MASK STRATEGY BASED ON OUR FRAMEWORK.

Source Dataset	Method	SFEW2.0	FER2013+	ExpW	AffectNet	JAFFE	Avg
RAFDB	Baseline(ResNet18)	45.18	60.23	50.93	41.63	50.70	49.73
	Baseline(MAE+ResNet18)	49.08	57.21	49.25	44.26	49.77	49.91
	LPL(CVPR'17) [23]	46.79	61.01	51.40	43.89	**56.81**	51.98
	ICID(Neuroc.'19) [14]	48.39	62.19	55.08	37.51	51.17	50.87
	PDEN(CVPR'21) [20]	41.28	52.02	35.33	38.43	53.52	44.12
	SNR(TMM'22) [15]	34.86	42.53	36.88	27.49	37.09	35.77
	Sequeener(NeurIPS'22) [37]	37.06	43.22	47.62	24.97	47.09	39.99
	EAC(ECCV'22) [47]	45.64	58.35	56.80	42.05	51.32	50.83
	MLA(ICMR'23)† [1]	51.38	61.28	56.19	44.49	54.46	53.56
	ActiveFER(ACII'23) [33]	43.12	57.54	44.46	44.06	50.00	47.82
	SADA(AAAI'23) [46]	49.40	59.86	51.58	39.11	49.60	49.91
	MARLIN(CVPR'23)‡ [4]	55.28	**73.01**	63.28	52.10	54.70	59.67
	Ours	**57.57**	72.85	**63.75**	**53.69**	55.75	**60.72**
	Method	SFEW2.0	RAFDB	FER2013+	ExpW	JAFFE	Avg
AffectNet	Baseline(ResNet18)	50.29	70.41	67.47	59.28	57.28	60.95
	Baseline(MAE+ResNet18)	54.13	75.13	70.65	60.74	53.52	62.83
	LPL(CVPR'17) [23]	51.75	72.13	68.59	60.43	58.22	62.22
	ICID(Neuroc.'19) [14]	50.64	69.33	69.46	59.83	53.52	60.56
	PDEN(CVPR'21) [20]	48.39	66.85	70.05	59.35	57.80	60.49
	SNR(TMM'22) [15]	47.25	69.46	67.73	58.85	51.17	58.89
	Sequeener(NeurIPS'22) [37]	52.35	72.42	70.46	61.13	58.22	62.92
	EAC(ECCV'22) [47]	47.02	69.81	67.95	59.65	53.95	59.68
	MLA(ICMR'23)† [1]	50.23	70.70	66.62	58.37	**60.44**	61.27
	ActiveFER(ACII'23) [33]	47.94	66.56	61.94	54.53	51.32	56.46
	SADA(AAAI'23) [46]	50.69	75.06	68.07	59.72	59.62	62.63
	MARLIN(CVPR'23)‡ [4]	54.88	80.37	74.95	63.19	58.09	66.42
	Ours	**55.05**	**80.93**	**75.11**	**64.83**	57.75	**66.73**

TABLE II

THE PERFORMANCE OF DIFFERENT ALGORITHMS ON THE TESTSET OF RAFDB. THE TRAINING SET IS RAFDB. THE BEST RESULT IS LABELED IN BOLD.

Method	Acc
ResNet18	85.50
Vit-base	87.22
MAE+ResNet18	91.07
DACL(WACV'21) [10]	87.78
APViT(TAFFC'22) [42]	91.98
EAC(ECCV'22) [47]	90.35
Ours	**92.01**

TABLE III

THE PERFORMANCE OF OUR ALGORITHM IN THE SCENARIO OF MULTI-SOURCE DOMAIN GENERALIZATION. THE BEST RESULTS ARE LABELED IN BOLD. THE SOURCE DATASETS ARE RAFDB AND SFEW.

Method	FER2013+	ExpW	AffectNet	JAFFE	Avg
Baseline (ResNet18)	62.46	52.92	41.23	53.87	52.97
Baseline (MAE+ResNet18)	60.19	51.79	45.74	56.34	53.52
MixStyle(ICLR'21) [49]	64.05	52.40	47.00	51.64	53.77
PDEN(CVPR'21) [20]	54.10	50.45	42.67	51.95	49.79
Sequeener(NeurIPS'22) [37]	55.54	50.28	46.86	50.85	48.88
MLA(ICMR'23)† [1]	63.18	52.24	46.72	54.15	54.07
SADA(AAAI'23) [46]	62.72	51.86	45.68	52.95	53.30
Ours	**68.15**	**62.42**	**53.91**	**59.62**	**61.03**

achieves an improvement of 7.51% over the related variant, i.e., ResNet18+MAE. Compared with other SOTA methods, our method achieves the best result in terms of average accuracy. Specifically, our method outperforms the SOTA method of SADA [46] by a margin of 7.73% in this multi-source domain scenario.

C. Ablation study

1) Ablation study on the proposed two modules: To study the performances of the proposed modules of ESGM, PDED, an ablation study is introduced in Table IV, where the RAFDB dataset is used as the source domain.

Compared with the baseline (first row of Table IV), replacing the random masking strategy with our ESGM (second row of Table IV) improves the average accuracy by 8.67%. Besides, the model with the module of PDED (third row of Table IV) gains improvements of 2.28%. When both ESGM and PDED are used, the accuracy is further improved to 60.72%. These results are probably due to that our ESGM achieves expression representation less sensitive to domain, and our PDED module enables the model to obtain more domain-invariant expression features conducive to CDFER.

2) Ablation study of the proposed decoupling strategy: To study the performance of the proposed decoupling strategy in PDED, we compare its performance with those of the baseline (without our PDED) and the other decoupling variants, as shown in Table V.

Table V shows that our PDED achieves a performance 2.14% higher than the baseline. In comparison with single decoupling, parallel decoupling and cyclic decoupling, our progressive decoupler outperforms them by the margins of 2.05%, 1.33% and 1.87%, respectively. The possible reason is that single decoupling does not fully decouple deep features. By contrast, cyclic decoupling contains more

TABLE IV

ABLATION STUDY OF THE PROPOSED MODULES OF ESGM AND PDED.

Source Dataset	ESGM	PDED	SFEW2.0	FER2013+	ExpW	AffectNet	JAFFE	Avg
RAFDB			49.08	57.21	49.25	44.26	49.77	49.91
	✓		54.36	70.46	61.46	52.40	54.21	58.58
		✓	53.67	58.59	49.90	44.31	54.46	52.19
	✓	✓	**57.57**	**72.85**	**63.75**	**53.69**	**55.75**	**60.72**

TABLE V

PERFORMANCES OF OUR DECOUPLING AND OTHER DECOUPLING METHOD. RAFDB IS USED FOR TRAINING. SINGLE DECOUPLING MEANS DECOUPLEING ONCE, PARALLEL DECOUPLING MEANS USING 4 PARALLEL DEOUCPERS TO DECOUPLE THE FEATURE, AND CYCLIC DECOUPLING MEANS TO DECOUPLE THE FEATURE FOR FOUR TIMES.

Method	SFEW2.0	FER2013+	ExpW	AffectNet	JAFFE	Avg
Without decoupling (baseline)	54.36	70.46	61.46	52.40	54.21	58.58
Single decoupling [5]	54.26	70.70	61.75	51.74	54.89	58.67
Parallel decoupling [35]	56.65	71.08	62.24	51.91	55.08	59.39
Cyclic decoupling [50]	55.14	70.62	61.89	52.28	54.31	58.85
Ours	**57.57**	**72.85**	**63.75**	**53.69**	**55.75**	**60.72**

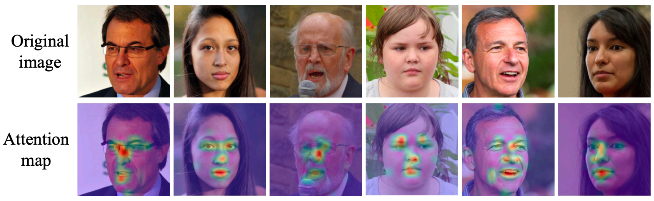

Fig. 4. The attention map of our ESGM result. The top row is the original images and the bottom row is the attention map A in Eq. (3) mapping to the original images. The regions with warmer color of the heatmap indicate the more attention by the attention block.

decoupling times, while the performance of this decoupling on the same feature is limited. By resorting to the parallel multi-branch decoupling, parallel decoupling outperforms the baseline by a margin of 0.81%, while it contains large redundancy among different branches, whose performance is still 1.33% lower than our PDED.

D. Algorithm Analysis

1) Representation visualization: For the attention map visualization, we use FFHQ as testset, and Fig. 4 shows that our attention map accurately focuses on expression-related areas, such as eyes, mouth, etc., allowing the network to reconstruct expression-aware areas and learn domain-invariant expression information.

To shed light on the capacity of our ESGM in reconstructing expression cues, we visualize the reconstructed expression images by random masking and our masking strategy in Fig. 5. To make the comparison of reconstruction visualization fair, we made the model trained with random masking adopt the same masking patches as our ESGM.

Fig. 5 shows that the restoration of expression details with random masking (MAE) is not clear enough, i.e. some blurring and distortion are produced in the reconstructed images, which may be because it can not mask expression-related regions. By contrast, our expression-aware masking produces much better reconstructed expressions, where blurring on the key expression-relevant regions is much reduced.

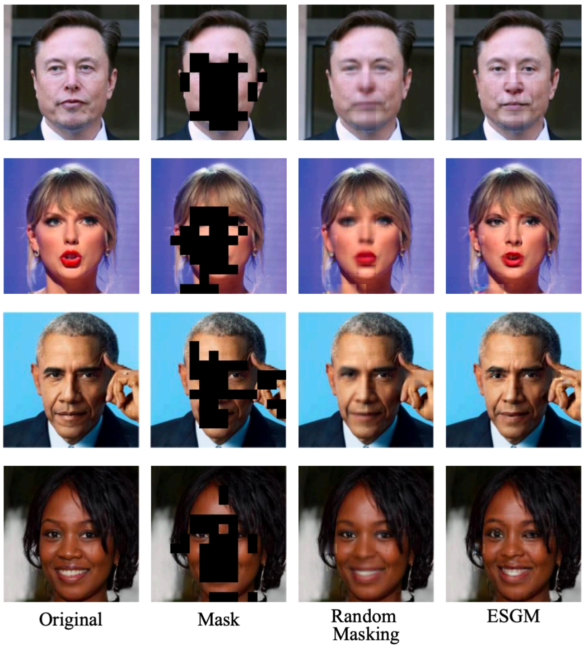

Fig. 5. The reconstruction visualization with random masking (MAE) and our ESGM. The 3rd column visualizes the results of the model trained with random masking, while the 4th column visualizes the results of the model trained with our strategy.

2) Decoupling Analysis: To study the performance of progressive decoupling for progressive separation of expression and domain features, we visualize the domain features $F_d^{(k)}$ extracted from each block of the ResNet via the t-SNE algorithm in Fig. 6, where the 2D distribution of features specific to the testing samples from RAFDB [21] and SFEW [8] is presented. It shows that the features in the shallow layers still appear intersecting parts between different domains, which are separated even further in the deeper block layer.

To further study the differences of the feature maps at

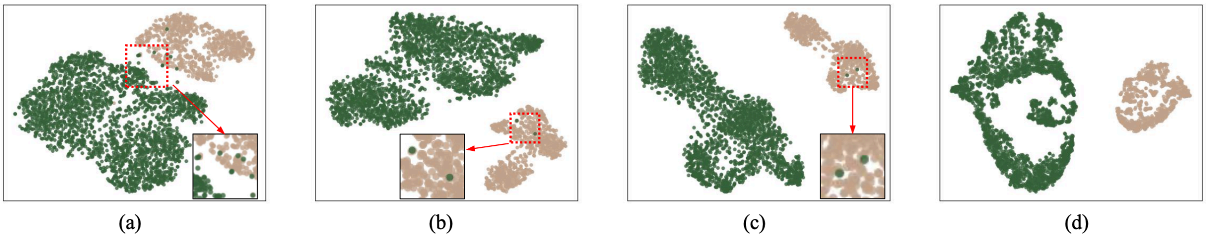

Fig. 6. The feature representation of examples from different datasets (RAFDB and SFEW) using the t-SNE algorithm. (a), (b), (c), (d) show the domain features separated from the four blocks of ResNet by decoupling, the features from the two domains gradually separate from each other. Green dots represent the domain features from RAFDB, while brown dots represent the domain features from SFEW.

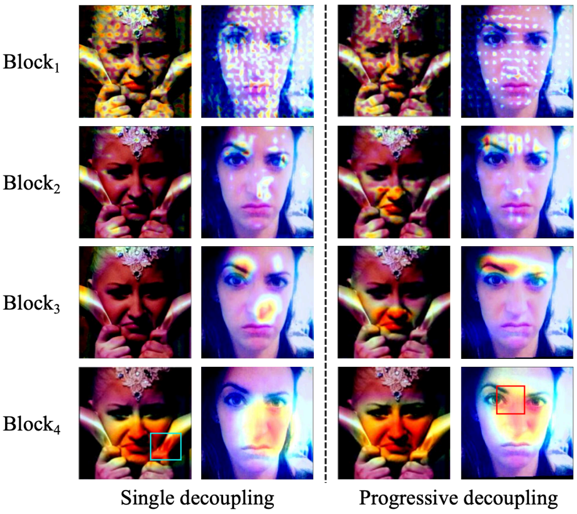

Fig. 7. The heatmap visualization of our progressive decoupling against the single decoupling by grad-cam [34]. The regions in red boxes indicate where our decoupling strategy outperforms the baseline in representing the expression-aware cues, while the regions in blue boxes indicate where the baseline decoupling focuses on the expression-insensitive regions.

different network blocks with single decoupling and our progressive decoupling, we use grad-cam [34] to visualize the heatmaps at different blocks by the two decoupling strategies in Fig. 7. It shows that our method not only reduces attention to expression-sensitive regions (blue boxes), but also increases attention to relevant regions (red boxes), compared to single decoupling. Meanwhile, our method can successfully pay attention to expression-related areas in shallow features.

3) Hyperparameter analysis: We present the sensitivity of our algorithm against mask ratio mr in Fig. 8. As the mask ratio increases, the average accuracy increases from 56.54% to the best of 60.72% and then decreases. These results show that an appropriate mask ratio is required to trade off the performances of reconstruction and expression representation. It is worth noting that the introduced masking and reconstruction in the pre-training stage are mainly used to recover expression information, thus, a mask ratio of 0.75 frequently used in existing works [11], [19], [24] is not used.

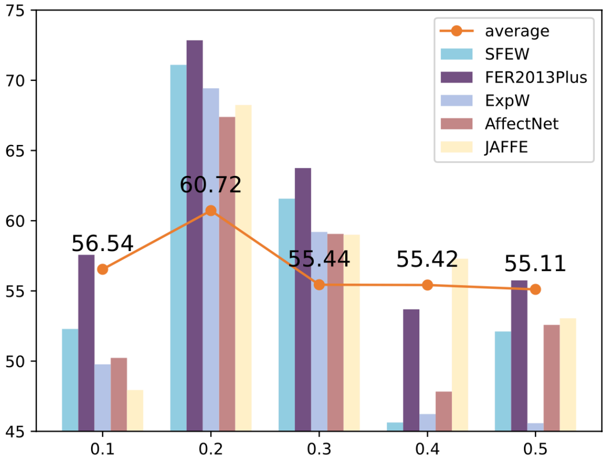

Fig. 8. The performance sensitivity against the mask ratio mr in Eq. (4). The model is trained on RAFDB, and tested on the other five datasets.

Fig. 5 shows that almost all expression cues has been masked with a mask ratio of 0.2. When the mask ratio is small, the network cannot learn a lot of expression-related knowledge. When the mask ratio is too high, the network pays too much attention to the other parts, thereby reducing the learning intensity of the expression-aware parts.

V. Conclusions and Discussions

To address the problems of low-efficiency representation learning, as well as entangled expression and domain features in cross-database facial expression recognition (CDFER), we propose expression-aware masking and progressive decoupler. Specifically, to make networks focus on expression cues during face masking and reconstruction, we propose an expression-aware masking strategy to occlude expression regions dynamically for representation learning. Meanwhile, to obtain domain-invariant expression features, we propose a progressive decoupler to separate domain and expression features by simultaneously taking into account the disentanglement performance in shallow layers. Extensive experiments on six public databases demonstrate that our approach outperforms state-of-the-art methods for CDFER in terms of generalization performance. Ablation studies and visualization results show the usefulness of each module. In our future work, expression-aware masking can be jointly learned with an AU recognition model, and specific masking strategies can be devised for different expression categories to further obtain a pre-trained encoder that is more robust to domain variation.

References

[1] A. Ballas and C. Diou. Cnns with multi-level attention for domain generalization. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, pages 592–596, 2023.

[2] H. Bao, L. Dong, S. Piao, and F. Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[3] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proc. ACM Int. Conf. Multimodal Interact.*, pages 279–283, 2016.

[4] Z. Cai, S. Ghosh, K. Stefanov, A. Dhall, J. Cai, H. Rezatofighi, R. Haffari, and M. Hayat. Marlin: Masked autoencoder for facial video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1493–1504, 2023.

[5] J. Chen, L. Yang, L. Tan, and R. Xu. Orthogonal channel attention-based multi-task learning for multi-view facial expression recognition. *Pattern Recognition*, 129:108753, 2022.

[6] J. Cheng, X. Mei, and M. Liu. Forecast-mae: Self-supervised pre-training for motion forecasting with masked autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8679–8689, 2023.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[8] A. Dhall, O. Ramana Murthy, R. Goecke, J. Joshi, and T. Gedeon. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proceedings of the ACM on international conference on multimodal interaction*, pages 423–426, 2015.

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[10] A. H. Farzaneh and X. Qi. Facial expression recognition in the wild via deep attentive center loss. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2402–2411, 2021.

[11] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[12] Z. Hou, X. Liu, Y. Cen, Y. Dong, H. Yang, C. Wang, and J. Tang. Graphmae: Self-supervised masked graph autoencoders. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 594–604, 2022.

[13] P.-Y. Huang, H. Xu, J. Li, A. Baevski, M. Auli, W. Galuba, F. Metze, and C. Feichtenhofer. Masked autoencoders that listen. *Advances in Neural Information Processing Systems*, 35:28708–28720, 2022.

[14] Y. Ji, Y. Hu, Y. Yang, F. Shen, and H. T. Shen. Cross-domain facial expression recognition via an intra-category common feature and inter-category distinction feature fusion network. *Neurocomputing*, 333:231–239, 2019.

[15] X. Jin, C. Lan, W. Zeng, and Z. Chen. Style normalization and restitution for domain generalization and adaptation. *IEEE Transactions on Multimedia*, 24:3636–3651, 2022.

[16] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[17] C.-H. Lee, Z. Liu, L. Wu, and P. Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5549–5558, 2020.

[18] 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*, pages 1536–1546, 2023.

[19] G. Li, H. Zheng, D. Liu, C. Wang, B. Su, and C. Zheng. Semmae: Semantic-guided masking for learning masked autoencoders. In *Advances in Neural Information Processing Systems*, volume 35, pages 14290–14302, 2022.

[20] L. Li, K. Gao, J. Cao, Z. Huang, Y. Weng, X. Mi, Z. Yu, X. Li, and B. Xia. Progressive domain expansion network for single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 224–233, 2021.

[21] S. Li and W. Deng. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing*, 28(1):356–370, 2019.

[22] S. Li and W. Deng. A deeper look at facial expression dataset bias. *IEEE Transactions on Affective Computing*, 13(2):881–893, 2022.

[23] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 2584–2593, 2017.

[24] X. Li, W. Wang, L. Yang, and J. Yang. Uniform masking: Enabling mae pre-training for pyramid-based vision transformers with locality. *arXiv preprint arXiv:2205.10063*, 2022.

[25] Y. Li, Z. Zhang, B. Chen, G. Lu, and D. Zhang. Deep margin-sensitive representation learning for cross-domain facial expression recognition. *IEEE Transactions on Multimedia*, 25:1359–1373, 2022.

[26] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

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

[28] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. Coding facial expressions with gabor wavelets. In *Proceedings Third IEEE international conference on automatic face and gesture recognition*, pages 200–205. IEEE, 1998.

[29] F. Ma, B. Sun, and S. Li. Facial expression recognition with visual transformers and attentional selective fusion. *IEEE Transactions on Affective Computing*, 14(2):1236–1248, 2021.

[30] C. Min, L. Xiao, D. Zhao, Y. Nie, and B. Dai. Occupancy-mae: Self-supervised pre-training large-scale lidar point clouds with masked occupancy autoencoders. *IEEE Transactions on Intelligent Vehicles*, 2023.

[31] A. Mollahosseini, B. Hasani, and M. H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affect. Comput.*, 10(1):18–31, 2017.

[32] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

[33] S. Roy and A. Etemad. Active learning with contrastive pre-training for facial expression recognition. In *International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8, 2023.

[34] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[35] W. Song, S. Shi, Y. Dong, and G. An. Heterogeneous spatio-temporal relation learning network for facial action unit detection. *Pattern Recognition Letters*, 164:268–275, 2022.

[36] L. Sun, Z. Lian, B. Liu, and J. Tao. Mae-dfer: Efficient masked autoencoder for self-supervised dynamic facial expression recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6110–6121, 2023.

[37] Y. Tatsunami and M. Taki. Sequencer: Deep lstm for image classification. In *Advances in Neural Information Processing Systems*, 2022.

[38] Z. Tong, Y. Song, J. Wang, and L. Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

[39] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, and L. Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010.

[40] C. Wang, J. Ding, H. Yan, and S. Shen. A prototype-oriented contrastive adaption network for cross-domain facial expression recognition. In *Proceedings of the Asian Conference on Computer Vision*, pages 4194–4210, 2022.

[41] Y. Xie, T. Chen, T. Pu, H. Wu, and L. Lin. Adversarial graph representation adaptation for cross-domain facial expression recognition. In *Proceedings of the 28th ACM international conference on Multimedia*, pages 1255–1264, 2020.

[42] F. Xue, Q. Wang, Z. Tan, Z. Ma, and G. Guo. Vision transformer with attentive pooling for robust facial expression recognition. *IEEE Transactions on Affective Computing*, 2022.

[43] K. Yan, W. Zheng, T. Zhang, Y. Zong, C. Tang, C. Lu, and Z. Cui. Cross-domain facial expression recognition based on transductive deep transfer learning. *IEEE Access*, 7:108906–108915, 2019.

[44] F. Yang, W. Xie, T. Zhong, J. Hu, and L. Shen. Augmented feature representation with parallel convolution for cross-domain facial expression recognition. In *Chinese Conference on Biometric Recognition*, pages 297–306. Springer, 2022.

[45] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[46] J. Zhang, H. Chao, A. Dhurandhar, P.-Y. Chen, A. Tajer, Y. Xu, and P. Yan. When neural networks fail to generalize? a model sensitivity perspective. In *AAAI Conference on Artificial Intelligence*, 2023.

[47] 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*, pages 418–434, 2022.

[48] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. From facial expression recognition to interpersonal relation prediction. *International Journal of Computer Vision*, 126(5):550–569, 2018.

[49] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang. Domain generalization with mixstyle. In *International Conference on Learning Representations*, 2021.

[50] X. Zou, Y. Yan, J.-H. Xue, S. Chen, and H. Wang. Learn-to-decompose: cascaded decomposition network for cross-domain few-shot facial expression recognition. In *European Conference on Computer Vision*, pages 683–700. Springer, 2022.