Generalization-Enhanced Feature-Wise and Correlation Attentions for Cross-Database Facial Expression Recognition

Weicheng Xie, *Member, IEEE*, Tao Zhong, Fan Yang, Siyang Song, Zitong Yu, and Linlin Shen, *Senior Member, IEEE*

***Abstract*—Cross-database facial expression recognition (CD-FER) has attracted increasing attention when evaluating the systems' generalization performance. Although the attention mechanism can capture the feature-wise importance or feature-correlation of expression sensitive regions, the attention-based network suffers from the overfitting to the source database, due to possible over-dependence on most salient features, without exploring feature characteristics during removal of feature redundancy. To address this issue, this paper introduces a multi-kernel competitive convolution in feature-wise attention to obtain more salient regions and let each kernel compete with others to enhance the expressive ability of features, by reducing attention overfitting to the source domain. For feature-correlation attention, we resort to a Monte Carlo-based dropout to not only reduce the over-learning of the feature relationship, but also model the dropout probability distribution more specifically by taking the characteristics of feature maps into account. Experimental results show that our algorithm achieves much better generalization performances than the state of the arts (SOTAs) on six publicly available datasets, in the scenarios of single source domain, multiple source domains and domain adaption.**

***Index Terms*—Facial expression recognition, cross database generalization, multi-kernel competitive convolution, Monte Carlo-based dropout.**

I. Introduction

FACIAL expression recognition (FER) and cross-database FER has received extensive attention in the field of computer vision due to its usefulness for various applications in human-computer interaction psychiatric health diagnosis and intelligent access control. Due to the possible domain gap (e.g. different distributions, races and poses) between databases, cross-database FER puts forward a higher requirement for the generalization capacity of the learned model.

Compared with the domain adaptation-based methods that alleviate the domain discrepancy between the source and target domains and thus require target domain data, domain generalization (DG) methods aim to train a model with enhanced generalization capacity without the aid of unseen target domain data. Specifically, typical DG methods involve data augmentation and generation [1], domain invariant feature representation [2] and meta-learning-based approaches [3].

For cross-database expression recognition [7], attention mechanism has been widely used for the feature-based or feature-correlation-based adaptive weighting [8], developing a feature representation common to different databases due to its adaptivity. However, attention-based networks frequently suffer from the overfitting to the source domain [9]. As shown in Fig. 1, both attention operations only focus on certain local regions, when they are trained on a single source domain and tested on the target domain. This means that the attention may over-fit to the source database. When the number of source domains increases, the well-trained network would focus on more sensitive regions of the entire face in a more comprehensive way. Thus, as shown in Fig. 1(e), we aim to achieve the effect of multi-source domains in the case of single source domain.

Although feature-wise attention (i.e., feature contributions are weighed adaptively in a feature map-wise manner) usually can make the network focus on certain local key information [8], it may not give enough attention to the contribution of sub-salient

Received 23 November 2024; accepted 5 January 2025. This work was supported in part by the Natural Science Foundation of China under Grant 62276170, Grant 82261138629, and Grant 62306061, in part by the Science and Technology Project of Guangdong Province under Grant 2023A1515011549 and Grant 2023A1515010688, in part by the Science and Technology Innovation Commission of Shenzhen under Grant JCYJ20220531101412030 and Grant JCYJ20210324094602007, in part by the Open Research Fund from Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ) under Grant GML-KF-24-11, and in part by the Guangdong Provincial Key Laboratory under Grant 2023B1212060076. *(Corresponding author: Linlin Shen.)*

Weicheng Xie is with the Computer Vision Institute, School of Computer Science & Software Engineering, Shenzhen University, Shenzhen 518060, China, also with the Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen 518132, China, also with the Guangdong Provincial Key Laboratory of Intelligent Information Processing, Shenzhen University, Shenzhen 518060, China, and also with the Department of Computer Science, University of Nottingham, Ningbo 315100, China.

Tao Zhong and Fan Yang are with the Computer Vision Institute, School of Computer Science & Software Engineering, Shenzhen University, Shenzhen 518060, China.

Siyang Song is with the School of Computer Science, University of Exeter, EX4 4SB Exeter, U.K.

Zitong Yu is with the Computer Science, Great Bay University, Dongguan 523000, China.

Linlin Shen is with the Computer Vision Institute, School of Computer Science & Software Engineering, Shenzhen University, Shenzhen 518060, China, also with the Guangdong Provincial Key Laboratory of Intelligent Information Processing, Shenzhen University, Shenzhen 518060, China, and also with the Department of Computer Science, University of Nottingham, Ningbo 315100, China (e-mail: llshen@szu.edu.cn).

Our codes are available at https://github.com/wuli-tao/paperCode.

This article has supplementary downloadable material available at https://doi.org/10.1109/TETCI.2025.3548727, provided by the authors.

Recommended for acceptance by N. Al Moubayed.

Digital Object Identifier 10.1109/TETCI.2025.3548727

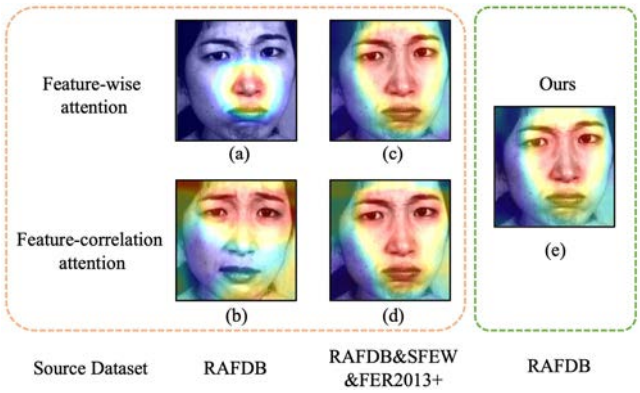


Fig. 1. Class Activation Map (CAM) [4] visualization of two attentions in two scenarios. (a) and (c) the focus of channel attention [5] for the target domain in the scenarios of the single-source domain and the multi-source domains, (b) and (d) the focus of feature-correlation-based self attention [6] for the target domain in the two scenarios, and (e) the focus of our method in the single-source domain scenario.

features [10], [11], while these features play an important role in robustness of the cross-domain recognition tasks [12]. Existing works mainly suppress salient regions, forcing the network to pay attention to sub-salient regions [13], or combine multi-scale global and local salient regions to improve feature representation learning [14]. For example, InceptionNet [15] uses multi-branch to obtain multi-scale information, SKNet [16] uses multiple kernels of different sizes to obtain features of different receptive fields, and EPSANet [17] resorts to multi-scale and channel-wise feature maps. Despite that these methods have shown the effectiveness of multi-convolution kernels in classification, the operations that direct concatenating features in InceptionNet [15] and treating different kernels equally important in SKNet [16], and considering only the channel-wise features in EPSANet [17], ignore the respective characteristic of each kernel.

In addition to the feature-wise attention, feature-correlation-based attention also has been frequently exploited for the contribution adaption of associated features, where the most typical representative is the vision transformer model [6]. Recent studies, e.g., [18] showed that there is a large amount of redundant attention weights in self-attention, resulting in over-learning to the source domain. Accordingly, several works aim to reduce this redundancy [19], and the most classic one is based on dropout [20]. Traditional dropout [20] discards neurons with a fixed probability, which has too much randomness, thus a number of dropout variants, e.g., blocks [21], channels [22] and attention [23], etc. are developed to reduce this randomness. Whereas, these methods learn redundant and less informative cues about the objects during the training [24]. Zeng et al. propose the CorrDrop [25], which samples an adaptive dropout mask in the Bernoulli distribution to discard the feature maps. However, the above-mentioned dropout are based on fixed distributions not capturing the specific characteristic of each sample, which thus do not make full use of the information specific to the feature maps.

To address the limitations discussed above, we propose a novel multi-kernel competitive convolution (MKCC) and Monte Carlo-based dropout for the cross-database FER task. (i) In contrast to existing single-kernel methods, our MKCC lets multiple kernels compete with each other to take into account both salient and sub-salient regions. To make use of the benefit of different kernels on behalf of the respective regions, intra- and inter- feature map cues specific to each kernel are aggregated via an attention mechanism. (ii) To take into account the specific characteristics of feature maps in reducing the co-adaptation among them, unlike the random or sample-unrelated manner widely used in existing works, our Monte Carlo-based dropout aims to integrate feature map-specific cues and reduces the feature-wise redundancy in a sample-dependent manner.

The contributions of this paper are summarized as follows

- A MKCC is proposed to adapt the feature-wise attention module by considering both sub-salient features and the most salient features, where a global sub-branch of kernel weighting can also make the model better generalizable to unseen data;
- A Monte Carlo-based dropout (MCD) is proposed to reduce the redundancy of global feature correlations. Compared with other dropout variants, our dropout is sample-dependent and well explores the specific characteristics of feature maps;
- Our method outperforms the related state of the arts for the task of cross-database FER in terms of single-source domain generalization, multi-source domain generalization and domain adaptation, and is a plug-and-play module directly integrable with other paradigms.

II. Related Work

A. Cross-Database Facial Expression Recognition (CDFER)

Methods of domain generalization were proposed to improve the recognition accuracy of the network for unseen databases in general fields, e.g., PDEN [26] utilized multiple sub-networks to simulate different domains, DIFEX [2] used high-level Fourier phases as domain invariant features, and SADA [27] resorted to suppressing model sensitivity in frequency space. Compared with these domain generalization tasks, cross-database generalization of FER is also rather challenging, due to the large differences among expressions from different databases, e.g., occlusion, poses, race and gender. Meanwhile, due to the limited scale of data, overfitting to the source domain exacerbates the difficulty of the problem. To this end, TDTLN [28] devised the cross-database-specific discriminative features, and Ji et al. proposed ICID [29] to learn both intra-category common features and inter-category discriminative features. Since these methods only use a single kernel to encode features, the sub-salient regions are not fully explored in the final feature representation. Ma et al. [30] used attention integration and Transformer structure to obtain global and local features. However, the overfitting of the attention mechanism to the source dataset is hardly studied and addressed.

In this work, to reduce over-learning on the source domain in CDFER, we resort to the multi-kernel competition mechanism to make model explore more on the sub-salient features, as well

as a self-attention with Monte Carlo-based Dropout to take into account the characteristics of feature maps.

B. Attention-Based Feature Representation

As an adaptive weighting mechanism, attention can be roughly categorized as the feature-wise and feature-correlation-based approaches [8]. For example, SENet [5] used channel attention, CBAM [31] combined channel and spatial attentions, SKNet [16] adaptively selected feature maps of different receptive fields, EPSANet [17] used multi-scale channel attention and Qu et al. [11] applied the attention mask to build the spatial attention. While these feature-wise attention methods mainly focus on the local information of the feature maps, such as channel or spatial information, they do not pay enough attention to the global cues, including those features of the feature maps that are not so salient.

Therefore, we design a multi-kernel competitive convolution module to explore this global information, i.e., the global sub-branch enlarges the weights of different sub-salient regions for robust recognition. In this way, there is a broader range of features for a network to explore, producing more generalizable feature representation helpful for cross-domain FER.

In contrast, feature-correlation-based attention builds up the dynamic correlation of features in the high-dimensional space, e.g., Transformer [32], the combination of CNN and Transformer [33], the Attentional Selective Fusion (ASF) of multiple features with the global-local attention [30]. However, these works do not consider the redundancy drawback of the feature correlations, i.e., the feature relationship for the source domain dataset is overfitted. Dropdim [34] drops part of the embedding dimensions of Transformer network to encourage it to encode more better-generalized features. Though this regularization can enhance the generalization capacity of the attention in Transformer, it does not fully explore the characteristics associated with the feature itself.

In this work, we resort to a feature-adaption dropout to reduce this redundancy in the attention mechanism.

C. Dropout-Based Generalization Improvement

Dropout [20] can make the model more generalizable and reduce overfitting, by setting the activation value of a certain neuron to zero with a certain probability. There have been many works [21], [22], [23], [35], [36] improving the original dropout.

Since the original dropout may destroy the overall structure of the feature map, Dropblock [21] was proposed to discard units in contiguous regions of the feature map in a structured manner. CamDrop [35] took into account the strength of the surrounding CAM to selectively discard some specific spatial regions, and DropKey [23] designed a dropout that sets Key as the dropout unit for the attention in Transformer. However, these methods mainly model the internal information of the feature maps, and do not pay sufficient attention on the co-adaptation or correlation among them.

To better reduce the co-adaptation between feature maps, Ding et al. [36] proposed Channel DropBlock (CDB), which clustered channels by a correlation matrix and randomly dropped groups of related channels. Xue et al. [33] proposed the Multi-Attention Dropping module to randomly drop a feature map with an uniform distribution. CorrDrop [25] sampled an adaptive mask in the Bernoulli distribution to discard the feature maps. However, the dropout masks obtained in these methods are based on the sampling of a fixed distribution, which cannot cope well with the distribution of different domains since each has its own characteristics.

Therefore, we propose a feature map-specific dropout with Monte Carlo-based probability distribution simulation. According to the characteristics of feature maps, our dropout is able to adaptively reduce the redundancy of feature correlation according to an instance-dependent dropout mask.

III. Methodology

In this section, we first illustrate the motivation of our newly-proposed modules or how they differ from the existing methods. Then, we introduce the main modules of our proposed framework, including Multi-Kernel Competitive Convolution (MKCC), Self-Attention with Monte Carlo-based Dropout (SAwMCD), and Inter-Scale Attention (ISA), as shown in Fig. 2. While MKCC aims to encode features from areas of different salient degrees, SAwMCD further reduces the redundancy in the global information between them to obtain a concise feature correlation representation. ISA is used to adaptively weigh feature maps with different scales for classification. Furthermore, we summarize the training process of the entire model and conduct an analysis of the proposed modules.

A. Motivation and Goals

1) Multi-Kernel Competitive Convolution (MKCC): For different databases, the feature distribution obtained by a single-convolution kernel will be overfitting to the specific distribution of each database, i.e., the convolution kernel is always learned to fit the distribution of the corresponding dataset. As a result, the well-trained model will focus on the areas that are specifically salient to images of the specific database, and may not sufficiently explore the remaining sub-salient areas that may informative for recognizing facial expressions collected by other datasets. Though improving the performance on the source domain, this characteristic is not beneficial for cross-domain FER.

As shown in Fig. 3, the setting of the single-kernel makes the network concentrate on few sensitive regions like the nose region, which may be not helpful for the unseen data with different expression cues. By contrast, multi-kernel setting makes the network explore different salient areas, such as eyes, nose, mouth, and eyebrow frown, allowing the network to better distinguish expressions from different databases. In addition, we allow these kernels to compete with each other, i.e., assigning a greater weight to the kernel with more discriminative ability.

2) Monte Carlo-Based Dropout: Traditional dropout inactivates a feature map based on a fixed dropout probability, as shown in Fig. 4(a). For the distribution-based dropout that drops a feature map based on a more general distribution, as shown

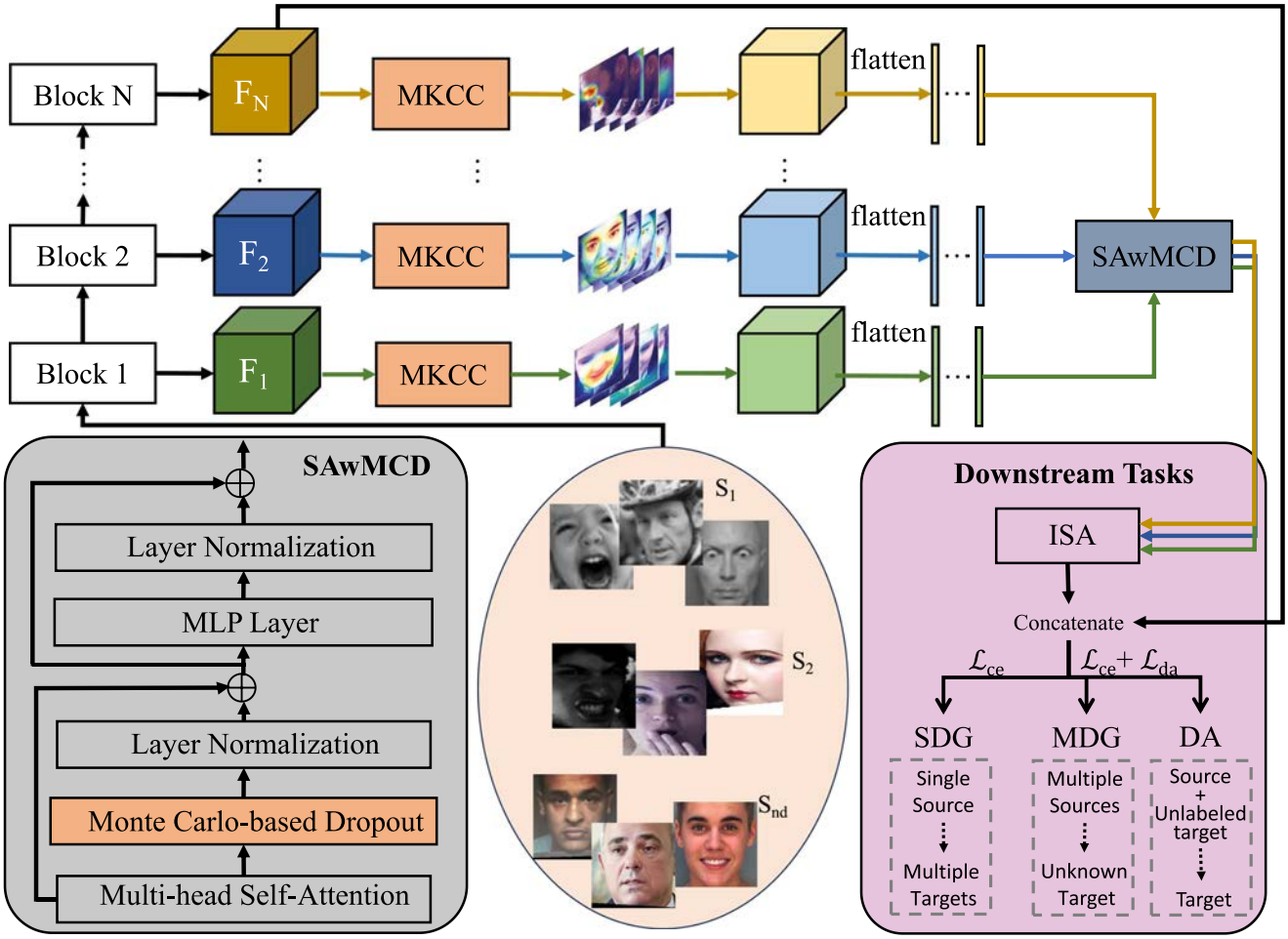


Fig. 2. Our method consists of the modules of Multi-Kernel Competitive Convolution (MKCC), Self-Attention with Monte Carlo-based dropout (SAwMCD) and Inter-Scale Attention (ISA). SDG, MDG and DA in the 'Downstream Tasks' block mean single-source domain generalization, multi-source domain generalization and domain adaptation, respectively, where the specific source and target domains are presented.

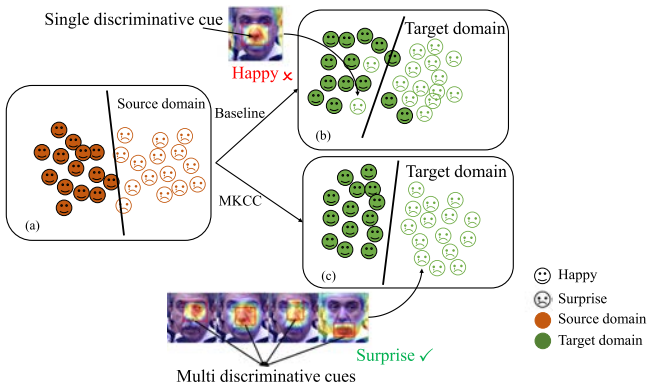


Fig. 3. The motivation of our MKCC. (a) The source domain representations of two expressions; (b) the representations of existing single-convolution kernel methods do not fully explore sub-salient regions; (c) the representations with our method leverages more diverse discriminative cues.

in Fig. 4(b), whether a feature map is dropped or not is determined by sampling a 0-1 mask value in a fixed distribution, e.g. Gaussian distribution. However, these methods use a dropout

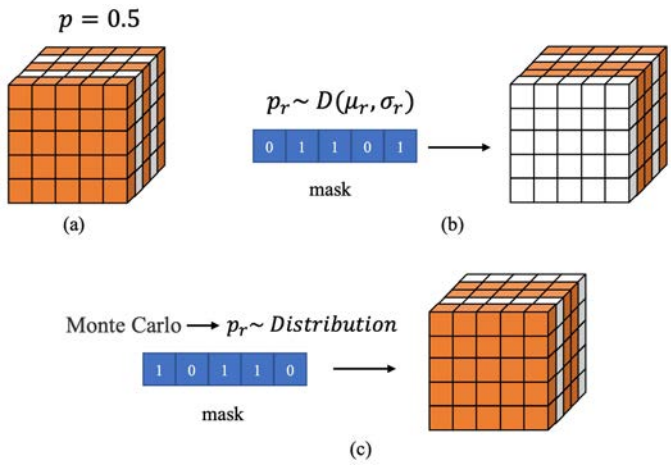


Fig. 4. Illustrations of different dropout variants. (a) Random dropout with a fixed probability, (b) sampling from a fixed probability distribution (e.g. normal distribution or uniform distribution) to generate a mask for dropout, (c) sampling from a varying distribution using the Monte Carlo method to generate a mask for dropout.

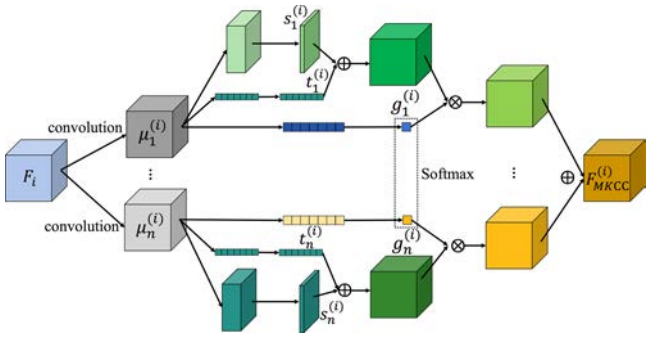


Fig. 5. Schematic diagram of our MKCC module.

probability distribution independent to the feature maps or the samples, thus do not take into account the characteristics of the feature maps themselves.

In this work, we propose a Monte Carlo-based dropout (MCD), as shown in Fig. 4(c), by using Monte Carlo sampling to randomly simulate the probability distribution of dropping a feature map. This approach dynamically adjusts the dropout probability based on the characteristics of the feature maps, making it sample-dependent. Specifically, a global average pooling (GAP) vector is used to capture the global characteristics of each feature map, which guides the Monte Carlo sampling process to generate adaptive dropout masks. Unlike traditional dropout methods that rely on fixed probabilities, MCD ensures that the dropout distribution reflects the specific characteristics of the feature maps, reducing over-learning to the source domain and improving generalization to unseen data. By exploring a broader range of features, including sub-salient regions, MCD enhances robustness in cross-database FER.

B. Multi-Kernel Competitive Convolution (MKCC)

To prevent the learned convolution kernel from overfitting to the source domain, we propose to use multiple kernels via different branches, where each branch contains a convolution and three attention operations that are learned to extract features of a unique degree of salience. The schematic diagram of our MKCC is shown in Fig. 5.

Specifically, for a batch of feature maps F_i, multiple branches are used to obtain feature maps of the same scale as:

$$\mu_k^{(i)} = Conv_k(Resize(F_i)), k = 1, \ldots, n \tag{1}$$

where F_i are the feature representation extracted from the i-th block, and are resized as $c \times h \times w$, where c, h, w are the numbers of channels, height and width of the feature map, respectively. $Conv_k$ means the k-th convolution operation and n is the number of its kernels.

To further enhance the feature representations corresponding to different salient regions, we resort to three attention strategies in the spatial, channel and global dimensions for each kernel-specific branch. First, in the spatial dimension, we integrate the information of different positions of each feature map as:

$$s_k^{(i)} = Sigmoid(Conv(ReLU(\mathbb{BN}(Conv(\mu_k^{(i)}))))) \tag{2}$$

where $\mathbb{BN}$ denotes the batch normalization [37]. Then, to model the importance of each channel, the channel-wise weight of the feature representation is obtained as:

$$t_k^{(i)} = Sigmoid(W_2(ReLU(W_1(\mathbb{GAP}(\mu_k^{(i)}))))) \tag{3}$$

where $\mathbb{GAP}$ represents the global average pooling and $W_1 \in \mathbb{R}^{\frac{c}{4} \times c}, W_2 \in \mathbb{R}^{c \times \frac{c}{4}}$ are the weights of the two fully connected (FC) layers.

Third, in addition to the local spatial and channel information, we obtain the specific global information with a global sub-branch to weight the importance of different kernels as follows:

$$g_k^{(i)} = Softmax(W_4(tanh(\mathbb{BN}(W_3(\mathbb{GAP}(\mu_k^{(i)})))))) \tag{4}$$

where $W_3 \in \mathbb{R}^{\frac{c}{4} \times c}, W_4 \in \mathbb{R}^{1 \times \frac{c}{4}}$ are the weights of the two FC layers, $\sum_{k=1}^{n} g_k^{(i)} = 1$. Based on this global sub-branch, each kernel is encouraged to compete with others, allowing network to assign a larger weight to the more important features.

Finally, the coefficients, i.e., $s_k^{(i)}$, $t_k^{(i)}$ and $g_k^{(i)}$ for the three attention sub-branches are assigned to the original feature map accordingly, and the feature map $F_{MKCC}^{(i)}$ after MKCC is formulated as follows

$$F_{MKCC}^{(i)} = \sum_{k=1}^{n} \mu_k^{(i)} \otimes g_k^{(i)} \otimes (s_k^{(i)} \oplus t_k^{(i)}) \tag{5}$$

where $\oplus$ represents the broadcasting addition and $\otimes$ means broadcasting element-wise multiplication [30].

C. Self-Attention With Monte Carlo-Based Dropout (SAwMCD)

To model the relationship between feature maps of different representations, i.e., $F_{MKCC}^{(i)}$ in (5) at the same scale, as well as obtain more diverse feature representations without large runtime overhead, we resort to the self-attention mechanism defined in the Transformer [32], i.e., this is achieved by optimizing the weights of learnable queries, keys and values. Specifically, we first use m FC layers to map $FF_{MKCC}^{(i)}$ into m specific representation spaces, where $FF_{MKCC}^{(i)} \in \mathbb{R}^{c \times hw}$ is the channel-wise token representation of the feature maps $F_{MKCC}^{(i)} \in \mathbb{R}^{c \times h \times w}$ (hw, c are the number and length of tokens) and each output is namely as a head, as shown in Fig. 2. Each head is formulated as:

$$head_j^{(i)} = Softmax\left(\frac{FF_{MKCC}^{(i)} W_j^Q (FF_{MKCC}^{(i)} W_j^K)^{\mathrm{T}}}{\sqrt{d}}\right) \times FF_{MKCC}^{(i)} W_j^V \tag{6}$$

where $W_j^Q, W_j^K, W_j^V \in \mathbb{R}^{c \times d}$ are the learnable weights for the queries, keys and values of the attention operation in the j-th head, respectively; $d = c/m$ is set to normalize the variance of each head output, and m denotes the number of heads.

Then, our MCD is performed on the concatenation of the heads specific to the i-th block, i.e., $H^{(i)} = [head_1^{(i)}, head_2^{(i)}, \ldots, head_m^{(i)}]$. As shown in Fig. 6, a global

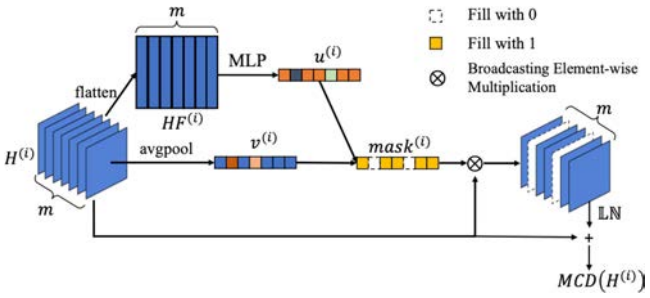


Fig. 6. The proposed Monte Carlo-based dropout (MCD).

average pooling operation is employed to obtain a mask that is more in line with the feature map in subsequent sampling for MCD. This is formulated as:

$$v^{(i)} = \mathbb{GAP}(H^{(i)}) \tag{7}$$

where $v^{(i)} \in \mathbb{R}^{m\times 1}$.

While the vector $v^{(i)}$ can encode the global information of each head, we further resort to the Monte Carlo sampling to randomly simulate the distribution information of the feature map, so as to make our dropout be more specific to the distribution of the feature map. Specifically, we formulate the thresholds of generating the binary mask for our dropout as:

$$u^{(i)} = HF^{(i)}\xi \tag{8}$$

where $u^{(i)} \in \mathbb{R}^{m\times 1}$, and the matrix $HF^{(i)} \in \mathbb{R}^{m\times(hw\times hw)}$ (m rows and $hw \times hw$ columns) denotes the flatten heads of $H^{(i)}$. $\xi \in \mathbb{R}^{(hw\times hw)\times 1}$ are sampled from the Kaiming initialization [38] distribution with the Monte Carlo technique, which is realized via a MLP. Since the parameters of MLP in the dropout method are not renewed with network back propagation, this MLP mapping is equivalent to randomly simulating a distribution containing the feature map-specific cues. Based on the threshold values $u^{(i)}$, we formulate the j-th dimension of the binary mask $mask^{(i)}$ as:

$$mask_j^{(i)} = \begin{cases} 0, & v_j^{(i)} > u_j^{(i)}, j = 1, 2, \ldots, m \\ 1, & otherwise \end{cases} \tag{9}$$

This mask can be adaptively adjusted according to the threshold $u_j^{(i)}$ that encodes the distribution of the feature map-specific cues via ξ in (8), and thus can represent the distribution of the feature map as well. Meanwhile, the sampling features that have larger responses than the average are masked out, reducing the possibility of overfitting to these salient features.

Finally, we perform dropout with the mask of heads to make the model better explore the sub-salient features, i.e., formulating the output of SAwMCD module as follows

$$\begin{cases} F_{MCD}^{(i)} = \mathbb{LN}(MCD(H^{(i)})W_{MLP}) + MCD(H^{(i)}) \\ MCD(H^{(i)}) = \mathbb{LN}(H^{(i)} \otimes mask^{(i)}) + H^{(i)} \end{cases} \tag{10}$$

where $\mathbb{LN}$ denotes the layer normalization, following Monte Carlo Dropout (MCD) and an additional MLP layer with the weight matrix of W_{MLP}, as shown in Fig. 2.

D. Training and Inference

1) Inter-Scale Attention (ISA): While the lower layers of the neural network learn the low-level general features, the higher layers tend to learn domain-specific features [39]. To make our model learn the domain-common features in lower layers in addition to the domain-specific features for the CDFER task, we integrate the cues from different blocks. Specifically, we formulate the weight of the feature representation with respect to the i-th block as follows

$$\omega^{(i)} = Sigmoid(ReLU(\mathbb{GAP}(F_{MCD}^{(i)})W_5)W_6) \tag{11}$$

where $\omega^{(i)} \in \mathbb{R}^{c\times 1}$, and W_5, W_6 are the weights of the two FC layers.

Finally, by integrating the cues of self-attention feature F_{ISA} and the feature F_N with the CNN backbone, the following feature representation F_{out} is used for classification

$$\begin{cases} F_{out} = Concat(F_{ISA}, F_N) \\ F_{ISA} = \sum_{i=1}^{N} F_{MCD}^{(i)} \otimes \omega^{(N)} \end{cases} \tag{12}$$

where $Concat(\cdot,\cdot)$ means stitching along the channel dimension and N indicates the number of blocks.

2) Overall Training: The single domain generalization is trained using the cross-entropy loss as follows

$$\mathscr{L}_{ce} = -\frac{1}{M_S}\sum_{e=1}^{M_S} y_e \log(\mathscr{F}(x_e, \theta_{\mathscr{F}})) \tag{13}$$

where M_S is the number of samples from the source domain data, x_e and y_e are the e-th training sample and its ground truth label, $\mathscr{F}$ means our employed network and $\theta_{\mathscr{F}}$ means its parameters.

For multi-domain generalization and domain adaptation, we utilize both cross-entropy loss $\mathscr{L}_{ce}$ and domain adversarial loss as follows

$$\mathscr{L}_{da} = -\sum_{q=1}^{nd}\frac{1}{M_{S_q}}\sum_{e=1}^{M_{S_q}} y_e^{S_q} \log(\mathscr{F}_{da}(x_e^{S_q}, \theta_{\mathscr{F}_{da}})) \tag{14}$$

where S_q represents the q-th of nd source domains and M_{S_q} represents the number of samples from it, $x_e^{S_q}$ and $y_e^{S_q}$ are the e-th training sample and its ground truth label from the q-th source domain. $\mathscr{F}_{da}$ means a domain adversarial network, i.e., three FC layers, and $\theta_{\mathscr{F}_{da}}$ means its parameters. For clarity, the training procedure of our algorithm is shown in Algorithm 1.

Lemma 1: The probability of dropping a head is approximate to the cumulative probability of a normal distribution.

Proof: Assume each of multi-layer perception (MLP) weight, i.e., the r-th dimension value $\xi_r \sim U(-a, a)$ obeys Kaiming initialization [38], where $a = \sqrt{\frac{6}{dim}}$ and dim is the dimension of MLP. Without loss of generality, let $O^{(j)} \in \mathbb{R}^{1\times(hw\times hw)}$ denote the j-th flatten head, i.e., the j-th row of $HF^{(i)}$ in (8). According to the central limit theorem:

$$\eta = u_j^{(i)} = \sum_{r=1}^{hw\times hw} O_r^{(j)}\xi_r \sim \mathscr{N}(0, \sqrt{b}\|O^{(j)}\|_2) \tag{15}$$

Algorithm 1: The Training Procedure of Our Method.

Input: Samples of source domains $\{x_e^{S_q}, y_e^{S_q}, q = 1, \ldots, nd\}$.

Output: Final model parameter $\theta_{\mathscr{F}}$ for the prediction.

1: **While** not converged **do**
2: Input feature maps F_i to the MKCC module for fusion feature map $F_{MKCC}^{(i)}$ in (5);
3: Input $F_{MKCC}^{(i)}$ to the SAwMCD module to get the redundant-reduced feature maps $F_{MCD}^{(i)}$ in (10);
4: Input F_{MCD} from all blocks into the ISA module to obtain multi-scale feature map representation F_{out} in (12);
5: Input F_{out} to the classifier and compute the classification loss $\mathscr{L}_{ce}$ in (13) (and additional domain adversarial loss $\mathscr{L}_{da}$ in the settings of multi-source and domain adaptation);
6: Update network parameters $\theta_{\mathscr{F}}$ using loss $\mathscr{L}_{ce}$ (and additionally update $\theta_{\mathscr{F}_{da}}$ using the loss $\mathscr{L}_{da}$ in (14) under the settings of multi-source and domain adaptation);
7: **end while**

where $b = \frac{[a-(-a)]^2}{12} = \frac{2}{dim}$ represents the variance of ξ_r, $O_r^{(j)}$ and $\|O^{(j)}\|_2$ represent the r-th value and the L_2-norm of the j-th flatten head.

Let Φ represent the distribution of η we simulated, φ is its probability density function. When η is greater than $mv = v_j^{(i)}$ in (7), i.e., $\frac{1}{hw \times hw} \sum_{r=1}^{hw \times hw} O_r^{(j)}$, it means to keep the r-th head, and the specific probability is formulated as follows

$$\Phi(mv) = \int \varphi(\eta > mv) d\eta = \int_{mv}^{+\infty} \varphi(\eta) d\eta \tag{16}$$

Thus, the probability of dropping this head is approximate to the cumulative probability of $\mathcal{N}(0, \sqrt{b}\|O^{(j)}\|_2)$. ■

IV. Experiment

A. Database and Experimental Setup

We evaluate the proposed approach on six public databases, i.e., RAF-DB [40], FER2013+[41], SFEW2.0 [42], AffectNet [43], ExpW [44] and JAFFE [45]. All the databases contain face images with seven expressions, i.e., six basic expressions and neutral.

RAF-DB database is a large-scale facial expression database, containing 29,672 in-the-wild facial images downloaded from the Internet. Each image has been independently labeled by about 40 annotators.

JAFFE database consists of 213 images from 10 different Japanese female subjects and the expressions were annotated by 60 annotators.

SFEW2.0 database was created by selecting static frames from the AFEW database [46]. It has been divided into three sets: 958 training samples, 436 validation samples and 372 test samples. In our experiment, the validation set is used for the testing.

TABLE I
Performances and Complexity Indicators in the Scenario of Single-Source Domain

Method \ Source Set	FLOPs(G)	Params(M)	SFEW2.0	RAFDB	FER2013+	ExpW	AffectNet
Baseline(ResNet18)	1.8	11.7	32.90	49.73	49.41	57.39	60.95
LPL(CVPR'17)	1.92	11.9	37.95	51.98	52.41	58.57	62.22
ICID(Neuroc.'19)	8.56	20.7	34.23	50.87	51.00	53.72	60.56
PDEN(CVPR'21)	1.85	47.56	35.17	44.12	51.78	56.72	60.49
SNR(TMM'22)	20.51	12.77	22.49	35.77	50.38	54.68	58.89
Sequeener(NeurIPS'22)	49.59	8.56	39.06	39.99	54.60	58.31	62.92
EAC(ECCV'22)	1.82	23.74	38.95	50.83	49.70	56.54	59.68
MLA(ICMR'23)	5.09	22.9	35.51	53.56	53.29	57.72	61.27
ActiveFER(ACII'23)	7.88	45.68	23.94	47.82	50.88	48.41	56.46
SADA(AAAI'23)	6.89	49.32	35.84	49.91	49.43	56.85	62.63
Ours	4.5	34.24	**39.14**	**53.86**	**54.99**	**58.81**	**63.57**

All the results are averaged over the remaining five databases and the best results are labeled in bold.

FER2013+ database annotations provide a set of new labels for the standard Emotion FER2013 [47] database. In FER2013+, each image has been labeled by 10 crowd-sourced taggers and it consists of 35,886 facial expression images with the size of 48× 48, including 28,708 training images, 3,589 public validation images and 3,589 private testing images. We use public validation and private testing images for the evaluation.

ExpW database contains 91,793 faces downloaded using Google image search.

Non-face images were removed in the label annotation process.

AffectNet database contains more than 1 M facial images collected from the Internet by querying three major search engines using 1,250 emotion related keywords in six different languages. We adopt facial images from 7 basic emotion categories for the experiments.

We use ResNet18 pre-trained on ImageNet as the backbone and the batch size is 32. For our MKCC, the input feature of each block is fixed to the size of $512 \times 14 \times 14$, and the number of kernels is set to 4. In our SAwMCD, the number of heads is set to 16. The networks are optimized via the AdaW algorithm with the learning rate of 0.0002, the weight decay of 0.01 and the momentum of 0.9. We train the networks for 50 epochs.

Specifically, for the single-source domain setting, we use the training set of the labeled source domain for the learning with the cross-entropy loss in (13), which is evaluated on the testing set of the target domain. For the multi-source domain setting, we leverage multiple labeled source domain training sets for training, with both the loss in (13) and domain adversarial loss in (14), and evaluate it on the target domain testing set. For the domain adaptation setting, we leverage the labeled source domain training set and unlabeled target domain training set for the training, with both the losses in (13) and (14), and evaluate it on the target domain testing set. The open source codes of the SOTA methods used for comparison are appended in the supplementary material.

B. Overall Performance Comparison

1) Single-Source Cross-Domain: We first evaluate the generalization performance and complexity indicators of our method trained on single source database, and the accuracy on the other databases are shown in Table I. We use RAFDB, FER2013+, SFEW, ExpW or AffectNet as source domain, and the remaining five databases as target domains for testing.

TABLE II
PERFORMANCES ON THE TESTING DATASET OF THE SOURCE DOMAIN IN THE SCENARIO OF SINGLE-SOURCE DOMAIN

Method \ Source Set	SFEW2.0	RAFDB	FER2013+	ExpW	AffectNet
Baseline(ResNet18)	39.45	85.50	83.52	69.88	57.91
Ours	**43.58**	**86.80**	**84.10**	**70.41**	**60.14**

TABLE III
PERFORMANCES IN THE SCENARIO OF MULTI-SOURCE DOMAIN

Method \ Source Set	RAF+FER	SFEW+FER	RAF+SFEW
Baseline	51.98	52.70	52.97
MixStyle(ICLR'21) [1]	52.21	55.78	53.77
PDEN(CVPR'21) [26]	45.53	46.03	49.79
Sequeener(NeurIPS'22) [49]	52.82	55.27	48.88
MLA(ICMR'23)† [48]	53.39	56.08	54.07
SADA(AAAI'23) [27]	53.21	55.91	53.30
Ours	**54.67**	**57.09**	**55.32**

RAF, FER, and SFEW are the abbreviations ofRAFDB, FER2013+, and SFEW2.0 databases, respectively. All the results are averaged over the remaining four databases and the best results are labeled in bold.

Compared with ICID [29] designed specifically for cross-domain FER, Table I shows that our method achieves an improvement of more than 3% on all source databases, and a large margin of 5.09% when ExpW is used as the source domain. This improvement maybe resulted from the multi-scale salient features obtained by our method, largely different from ICID that uses only a single feature in the last layer. Compared with the generalization-oriented multi-level attention, i.e. MLA [48] that is designed for FER, our algorithm outperforms it by 1% on all target domains. The achieved better generalization performance may be due to the alleviated over-learning in the self-attention operation via the introduced MCD. Compared with other state-of-the-art (SOTA) DG methods, including PDEN [26], Sequeener [49], SADA [27], 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 13.87% over the Sequeener [49].

Meanwhile, our model achieves a better balance between efficiency and effectiveness, with a computational complexity of 4.5 M that is largely lower than the complex models like SNR (20.51 M) and Sequeener (49.59 M), as well as a parameter count of 34.24 G, that is largely lower than those of current state of the arts, e.g. SADA (49.32 G). Despite its moderate resource overhead, it achieves the state-of-the-art performances.

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 the baseline.

In addition to the SOTA performances on the target domains in Table I, Table II shows that our method also achieves improvements over the baseline on the testing datasets of all source domains, e.g., our method achieves the highest improvement of 4.13% on the test dataset of SFEW. There is also a 2.23% improvement on the challenging database AffectNet in the wild.

2) Multi-Source Cross-Domain: We further evaluate our method of domain generalization in the setting of multiple source domains, and Table III presents the results.

Compared with the setting of single source domain of SFEW, the average accuracy on the target domain database has significantly improved by a margin of 16.18% when the domains of SFEW and RAFDB are used for the training. Compared with other SOTA methods, our method achieves the best result in terms of average accuracy. Specifically, our method outperforms the Sequeener [49] by a margin of 6.44% in the multi-source domain scenario and MixStyle [1] by 1.55% when RAFDB and SFEW2.0 are used as source domains. Compared with MLA [48], which also extends to the scenario of multi-source domains, we achieve improvements larger than 1% in all cases.

3) Domain Adaptation: To study the performance of our algorithm in terms of domain adaptation, we present the results of the SOTA methods [50], [51], [52], [53], [54] and ours in Table IV, where ECAN [51] is specially designed for cross-domain expression recognition.

Table IV shows that our method better trades off the domain-adaption performance on each of target dataset and achieves the best average results on all databases, no mater when the large-scale dataset, i.e., AffectNet or the small-scale dataset, i.e., SFEW2.0 is used as the source dataset. By contrast, other SOTAs such as JUMBOT [52] do not perform stably. Taking JAFFE as the target dataset for example, while JUMBOT [52] achieves the SOTA performance, i.e., 65.79% when AffectNet is used as source dataset, its performance declines to only 19.74% when the small-scale dataset SFEW2.0 is used for training. These results show that our method can also work well for the laboratory-controlled database JAFFE that has a large domain shift in terms of gender and race.

Since EADA [57] and ELS [58] are designed specially for the domain adaptation task, they utilize additional strategies that differ much from ours, such as data selection, data augmentation, and other paradigms. To evaluate the performance of our algorithm generalized onto other paradigms, we use them as the baselines, and integrate our modules with their methods and present the results in Table VI. One can see that our module outperforms the baselines by 1%-2% in accuracy, i.e., our modules are plug-and-play and can effectively enhance the recognition ability of the SOTA models.

4) Dropout Method: Our proposed method builds upon Monte Carlo Dropout (MCD) and leverages its uncertainty estimation capabilities, which is especially advantageous for facial expression recognition in challenging conditions. We further compare our dropout with several dropout algorithms (reproduced the codes by ourselves), i.e. CDB [36], MC-FerNet [55] and RIFAD [56] in the scenario of single-source domain (with the same setting of Table I) and present the results in Table V. As shown in Table V, by combining MCD with techniques specifically tailored for facial expression data, our approach delivers the best performance, surpassing other dropout methods.

C. Ablation Study

In this section, we perform ablation study on each module of our method.

1) Ablation Study on Three Modules: To validate the proposed modules of MKCC, SAwMCD, and ISA, an ablation study

TABLE IV
PERFORMANCES IN THE SCENARIO OF CROSS-DOMAIN ADAPTATION

Method	Source set	SFEW2.0	RAFDB	FER2013+	ExpW	AffectNet	JAFFE	Avg
Baseline	SFEW2.0	-	40.48	46.72	36.45	26.63	28.26	35.71
SWD(CVPR'19) [50]	SFEW2.0	-	**45.08**	38.35	37.53	29.01	27.63	35.52
ECAN(TAFFC'20) [51]	SFEW2.0	-	27.02	38.33	32.02	27.83	25.00	30.04
JUMBOT(ICML'21) [52]	SFEW2.0	-	32.43	41.67	34.52	28.71	19.74	31.41
AngularGap(MM'22) [53]	SFEW2.0	-	42.01	47.42	31.79	23.34	29.71	34.85
SRoUDA(AAAI'23) [54]	SFEW2.0	-	33.16	46.72	**37.93**	24.29	23.16	33.05
Ours	SFEW2.0	-	42.10	**49.80**	37.20	**29.46**	**30.26**	**37.76**
Baseline	AffectNet	44.10	66.38	64.36	57.05	-	48.68	56.11
SWD(CVPR'19) [50]	AffectNet	**50.42**	69.38	66.79	57.51	-	56.99	60.22
ECAN(TAFFC'20) [51]	AffectNet	47.71	67.49	63.75	59.57	-	53.62	58.43
JUMBOT(ICML'21) [52]	AffectNet	44.95	61.83	58.93	49.80	-	**65.79**	56.26
AngularGap(MM'22) [53]	AffectNet	46.79	67.89	65.97	58.16	-	49.19	57.60
SRoUDA(AAAI'23) [54]	AffectNet	49.72	67.16	**66.82**	58.43	-	53.26	59.08
Ours	AffectNet	48.71	**70.57**	65.94	**60.75**	-	55.95	**60.38**

The best results are labeled in bold.

TABLE V
COMPARISON WITH RECENT DROPOUT-BASED METHODS

Method \ Source Set	SFEW2.0	RAFDB	FER2013+	ExpW	AffectNet
CDB (BMVC'21) [36]	33.21	42.36	52.23	53.51	59.24
MC-FerNet (ACAIT'23) [55]	36.25	51.89	53.22	54.59	60.25
RIFAD (CVIU'23) [56]	38.26	53.26	51.24	56.37	62.84
Ours	**39.14**	**53.86**	**54.99**	**58.81**	**63.57**

All the results are averaged over the remaining five databases and the best results are labeled in bold.

TABLE VI
RESULTS OF OUR ALGORITHM IN TERMS OF CROSS-DOMAIN ADAPTATION WHEN A SOTA ALGORITHM IS USED AS THE BASELINE

Method \ Source Set	SFEW2.0	RAFDB	FER2013+	ExpW	AffectNet
EADA(AAAI'22) [57]	49.50	57.00	58.99	61.64	64.18
EADA+Ours	**50.04**	**59.91**	**60.14**	**62.19**	**64.88**
ELS(ICLR'23) [58]	33.62	49.88	54.30	56.14	59.08
ELS+Ours	**35.86**	**51.71**	**56.51**	**56.87**	**60.02**

All the results are averaged over the remaining five databases and the best results are labeled in bold.

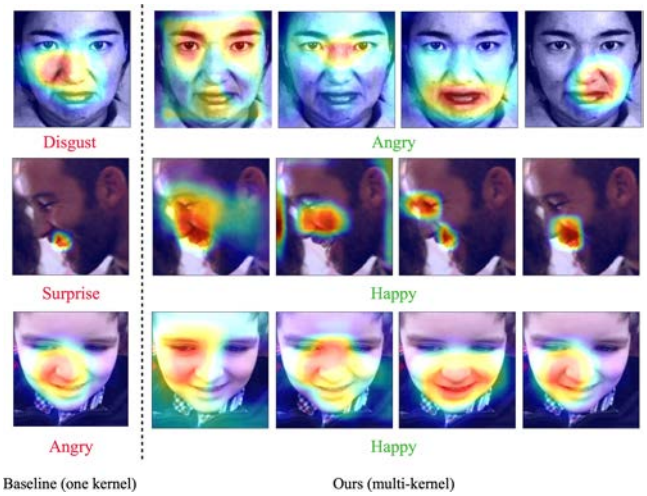


Fig. 7. The visualization of the learned feature maps. The left column shows those of the baseline, the right four columns present those of the respective convolution kernels in MKCC. The red and green represent the wrongly and correctly classified categories, respectively.

is introduced in Table VII, where RAFDB or AffectNet is used as the source domain.

Compared with the baseline, the MKCC module improves the average accuracy by (2.51%, 0.85%) and gains improvements of (0.89%, 0.94%) with the addition of SAwMCD module. This shows that the multiple levels of salient features with MKCC and the reduction of feature map correlation with SAwMCD are both helpful to cross-domain FER. The proposed ISA further improves the accuracy by the integration of multiple-layer features with different scales, which facilitates our algorithm to achieve the SOTA performances of (53.86%, 63.57%).

2) Ablation Study on the MKCC Module: For this ablation study, we compare the results of multi-kernel representative SKNet [16], EPSANet [17] and ours in Table VIII.

It shows that our MKCC achieves the average accuracy of 53.86%, which is 1.99% higher than that with SKNet and 1.69% higher than that with EPSANet. Instead of assigning the same weights for all the branches in SKNet, our global sub-branch in MKCC enables the network to assign specific weight to each scale of features in the corresponding branch by encouraging competition between different kernels, can thus better cope with samples from different domains. When this global sub-branch is imposed on SKNet, it gains an improvement of 1.2%. Compared with EPSANet which only uses multi-scale channel-wise features, our algorithm additionally considers the spatial and global features, and achieves an improvement of 1.69%.

To shed light on how the proposed MKCC works, we visualize the feature maps obtained by the baseline and our MKCC in Fig. 7. It visualizes the feature maps learned by the baseline model and the proposed MKCC module. The baseline model focuses primarily on the most salient regions (e.g., the nose), which may not generalize well to unseen data. In contrast, the MKCC module employs multiple convolution kernels to capture features from regions with varying degrees of salience, such as the eyes, mouth, and eyebrow frown. For the reasons, sine the output of each kernel is weighted based on its discriminative ability, making the kernels compete dynamically. This allows the model to explore a broader range of features, including sub-salient regions, to facilitate recognizing expressions across different databases. By leveraging multiple kernels and attention mechanisms, MKCC enhances the model's ability to generalize to unseen data, as demonstrated in the visualization.

TABLE VII
ABLATION STUDY OF THE PROPOSED MODULES OF MKCC, SAWMCD AND ISA

Source Dataset	MKCC	SAwMCD	ISA	SFEW2.0	FER2013+	ExpW	AffectNet	JAFFE	Avg
RAFDB				45.18	60.23	50.93	41.63	50.70	49.73
	✔			49.62	60.54	**52.45**	44.09	54.52	52.24
		✔		49.87	61.26	51.13	43.93	53.58	51.95
			✔	49.00	60.81	51.25	43.86	53.05	51.49
	✔	✔		51.71	62.02	51.74	44.11	56.05	53.13
	✔	✔	✔	**52.75**	**62.08**	52.17	**44.54**	**57.75**	**53.86**
Source Dataset	MKCC	SAwMCD	ISA	SFEW2.0	RAFDB	FER2013+	ExpW	JAFFE	Avg
AffectNet				50.29	70.41	67.47	59.28	57.28	60.95
	✔			51.06	70.85	69.13	59.72	58.22	61.80
		✔		**51.98**	70.53	69.42	59.91	58.81	62.13
			✔	51.52	69.39	69.77	59.61	58.62	61.78
	✔	✔		50.89	71.69	70.29	60.87	59.95	62.74
	✔	✔	✔	51.88	**72.59**	**71.63**	**61.19**	**60.56**	**63.57**

TABLE VIII
PERFORMANCES OF SKNET [16] AND OUR MKCC, 'GLOBAL' MEANS OUR GLOBAL SUB-BRANCH IN (4)

Method \ Source Set	SFEW2.0	FER2013+	ExpW	AffectNet	JAFFE	Avg
Ours-MKCC+SKNet [16]	47.25	61.34	51.21	43.11	56.42	51.87
Ours-MKCC+SKNet [16]+global	49.75	**62.11**	52.09	44.14	57.28	53.07
Ours-MKCC+EPSANet [17]	49.79	60.75	51.47	43.51	55.32	52.17
Ours	**52.75**	62.08	**52.17**	**44.54**	**57.75**	**53.86**

RAFDB is used for training.

TABLE IX
PERFORMANCES OF DROPOUT VARIANTS WITH DIFFERENT STRATEGIES OF MASK GENERATION

Method \ Source Set	SFEW2.0	FER2013+	ExpW	AffectNet	JAFFE	Avg
Baseline	50.06	60.45	51.20	42.46	53.52	51.54
Bernoulli(0.5) [59]	46.56	60.01	49.52	43.46	55.87	51.08
$\mathcal{N}(0.5, \frac{1}{2})$ [59]	44.95	61.61	51.36	**44.89**	56.81	51.92
$U(0,1)$ [59]	48.39	59.19	51.14	43.23	54.52	51.30
$\mathcal{N}(0.5, \frac{1}{2\sqrt{3}})$	49.64	59.88	50.48	43.89	56.34	52.05
Mask inversion	47.25	60.09	49.95	42.46	54.46	50.84
$\mathcal{N}(0, \sqrt{b}\|O^{(j)}\|_2)$	50.25	61.78	50.11	43.31	56.81	52.45
DynamicViT [19]	50.31	59.36	51.71	44.31	55.52	52.24
Ours (MCD)	**52.75**	**62.08**	**52.17**	44.54	**57.75**	**53.86**

Baseline denotes our method without the proposed dropout. Bernoulli distribution *Bernoulli*(0.5) [59], Gaussian distribution $\mathcal{N}(0.5, \frac{1}{2})$ [59] and $\mathcal{N}(0.5, \frac{1}{2\sqrt{3}})$ (the setting when mv is assumed to be a random variable obeying the same distribution as η in Eq. (16), please refer to the appendix material for the details), and the uniform distribution $U(0,1)$ [59] is used to simulate ξ in Eq. (8) for the ablation study.Meanwhile, the asymptotic distribution of η in Eq. (15), i.e., $\mathcal{N}(0, \sqrt{b}\|O^{(j)}\|_2)$ is used to surrogate the proposed Monte Carlo sampling for the evaluation. DynamicViT [19] adjusts the dropout mask with gradient backpropagation. Mask inversion inverts the values of 0 and 1 in the mask generation in Eq. (9) during the training on RAFDB.

3) Ablation Study on the SAwMCD Module: To study the performance of the proposed MCD, we compare its performance with those of the baseline (without our MCD) and the dropout variants with different mask samplings and updating strategies in Table IX. For this comparison, DynamicViT [19] that uses Gumbel Softmax method to dynamically adjust the dropout mask via gradient backpropagation, is also adopted.

Table IX shows that our MCD achieves a performance 2.32% higher than the baseline. In comparison with mask dropout based on fixed distributions, our Monte Carlo-based strategy outperforms the Bernoulli, Gaussian, and uniform distributions by the margins of 2.78%, 1.94% and 2.56%, respectively, which means that our MCD can better reflect the specific characteristics of features than those with hand-crafted or feature-independent distributions. Compared with the settings of $\mathcal{N}(0.5, \frac{1}{2\sqrt{3}})$ and $\mathcal{N}(0, \sqrt{b}\|O^{(j)}\|_2)$, our method leverages the Monte Carlo sampling to match the distribution of the original feature maps more closely, can thus make the dropout sample-adaptive. Compared with the strategy of reversing the mask values, our proposed dropout also improves the performance significantly. This is probably because the mask reversing will encourage the network to focus more on the neurons with larger responses, which lead to over-learning phenomenon of the model. Compared with the DynamicViT [19] that dynamically updates the dropout mask, our MCD can still achieve an improvement of 1.62%.

TABLE X
PERFORMANCES OF DIFFERENT ATTENTION VARIANTS TRAINED ON RAFDB

Method \ Source Set	SFEW2.0	FER2013+	ExpW	AffectNet	JAFFE	Avg
MKCC&SAwMCD	47.71	58.69	50.72	43.54	55.40	51.21
MKCC in SAwMCD	49.63	60.45	51.21	44.12	55.34	52.15
SAwMCD+MKCC	48.17	60.64	**53.01**	44.20	54.93	52.19
MKCC+SAwMCD (Ours)	**52.75**	**62.08**	52.17	**44.54**	**57.75**	**53.86**

MKCC in SAwMCD means using MKCC and SAwMCD alternately. MKCC&SAwMCD means using both MKCC and SAwMCD in the feature-wise attention. SAwMCD+MKCC means applying MKCC after SAwMCD, i.e., the default order of MKCC and SAwMCD is swapped.

In order to shed light on the role of SAwMCD in alleviating the over-learning, we visualized the cosine similarity between tokens, i.e. $FF^{(i)}_{MKCC}$ in (5), under the conditions of non-dropout, original dropout and SAwMCD. As shown in Fig. 8, the similarity between the tokens of the original attention with the setting of non-dropout is relatively higher, implying larger possibility of over-learning. Compared with the original dropout that can reduce the co-adaptation between tokens in original attention to some extent, our method further reduces the redundancy between tokens, thereby largely reducing the over-learning phenomenon of the learned model.

4) Ablation Study on the Attention Variants: To study the performances of our MKCC and SAwMCD on the feature-wise and feature-correlation attentions, we also introduce different attention variants and present their performances in Table X. For the setting of 'MKCC in SAwMCD', the MKCC is added to each SAwMCD layer, and MKCC and SAwMCD are used alternately. For the setting of MKCC&SAwMCD, the feature maps after the weighting of the three attention sub-branches in MKCC using (2)–(4), are input to SAwMCD, and the feature maps from the branches of all the kernels are finally added up.

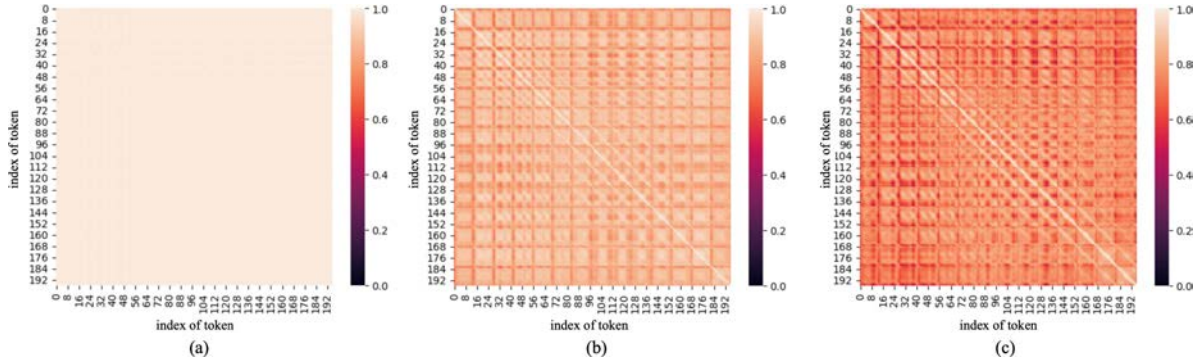


Fig. 8. Heatmaps of the cosine similarity between different tokens (i.e. $FF^{(i)}_{MKCC}$) of non-dropout (a), original dropout (b), and SAwMCD (c) on RAFDB.
.

For the setting of SAwMCD+MKCC, the output feature maps of SAwMCD are used as the input of MKCC. Schematic diagrams of the above attention variants are shown in the supplementary material.

Table X shows that the average accuracy declines from 53.86% to 52.19% when the order of MKCC and SAwMCD is reversed, i.e., SAwMCD is used to dropout the feature maps and MKCC is used in aggregating the cues from multiple heads. When alternatively employing the modules of MKCC and SAwMCD, one can see that the average accuracy drops by 1.71%, compared with the proposed setting. When using MKCC and SAwMCD together in the feature-wise attention, one can see that the average accuracy drops by 2.65%.

To speculate the performance drop in these module settings, we argue that the network is apt to focus on the most salient parts in the feature-wise representation, thus, the sub-salient features captured with our MKCC are preferred. By contrast, the feature correlation redundancy is easily to be induced when high-dimensional feature-to-feature relationship substitutes for the feature-wise representation, our proposed SAwMCD is thus more preferred.

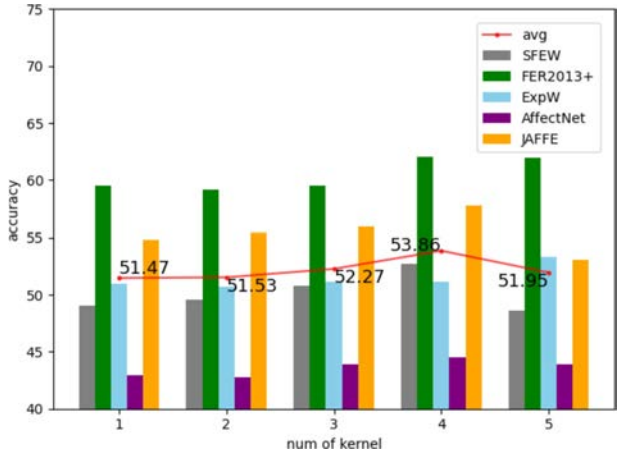


Fig. 9. The performance sensitivity against the number of kernels, i.e., n in (1). The models are trained on RAF-DB, and tested on the other five databases.

D. Algorithm Analysis

In this section, we first investigate the sensitivity of the performance of our algorithm against hyperparameter settings, and then visualize the learned features using the CAM [4] and the t-SNE technique [60].

1) Hyperparameter Analysis: We present the sensitivity of our algorithm against the number of kernels, i.e., n of (1) in Fig. 9. As the number of kernels increases, the average accuracy increases from 51.47% to the best of 53.86% and then degrades. These results show that an appropriate number of kernels is demanded to trade off the feature diversity of the representation and the capacity of these features in generalization ability enhancement.

2) Representation Visualization: To study the performance of our algorithm on each expression, we show the confusion matrix of the baseline and our method in Fig. 10. For the same category of expression in different databases, the recognition accuracy could differ much. When JAFFE or ExpW is used as the target domain, our algorithm achieves an accuracy of 73% for 'sad' on JAFFE, while only 33% on ExpW. This shows that the same expression may show large variation in different databases, which reflects the usefulness of the proposed multi-kernel to explore more sub-salient regions to cover different data. In addition, compared with the baseline, our method has obvious advantage in telling apart difficult expression categories, such as 'fear' and 'disgust'. For the 'fear' expression in FER2013+, the accuracy of the baseline is only 4%, but our method can reach 25%.

For the visualization of 2D feature representations with t-SNE [60], we demonstrate those learned by the baseline and our model in Fig. 11. As shown in the red boxes of Fig. 11, there are large overlaps among features of the baseline model for different categories. By contrast, this overlap is largely reduced by our method. Meanwhile, our method separates these features much better than the baseline.

For the feature map visualization with CAM [4], we use RAFDB as the source domain, and the databases of JAFFE and SFEW as the target domain, and show the results in Fig. 12. One can see that the attention of the baseline mainly focuses on limited expression-sensitive regions, while our method pays

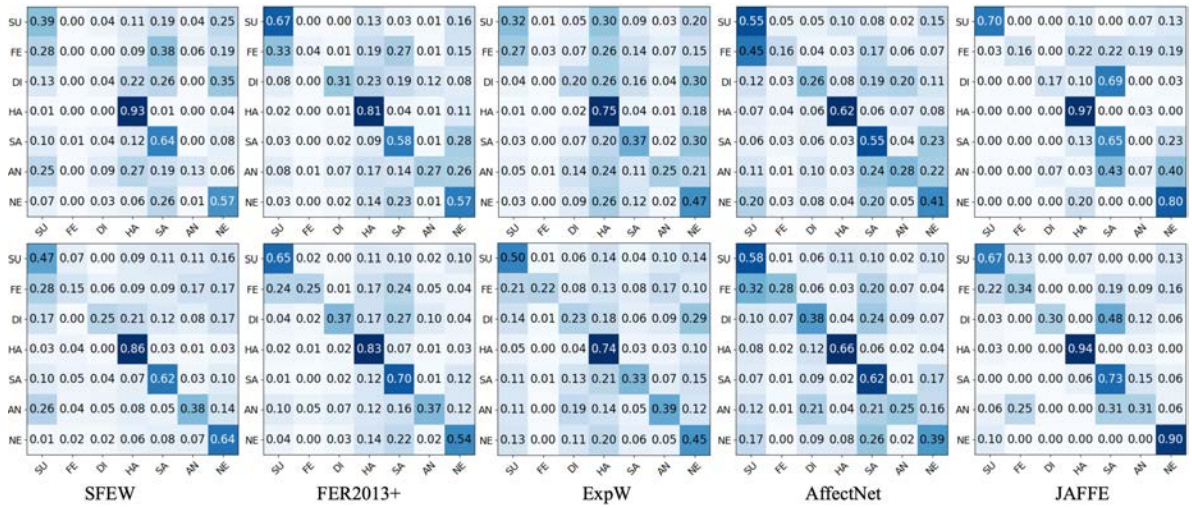


Fig. 10. Confusion matrices with the baseline (1st row) and ours (2nd row) on five databases. AN, DI, FE, HA, NE, SA and SU are the abbreviations of angry, disgust, fear, happy, neutral, sad and surprise, respectively. RAFDB is used for training.

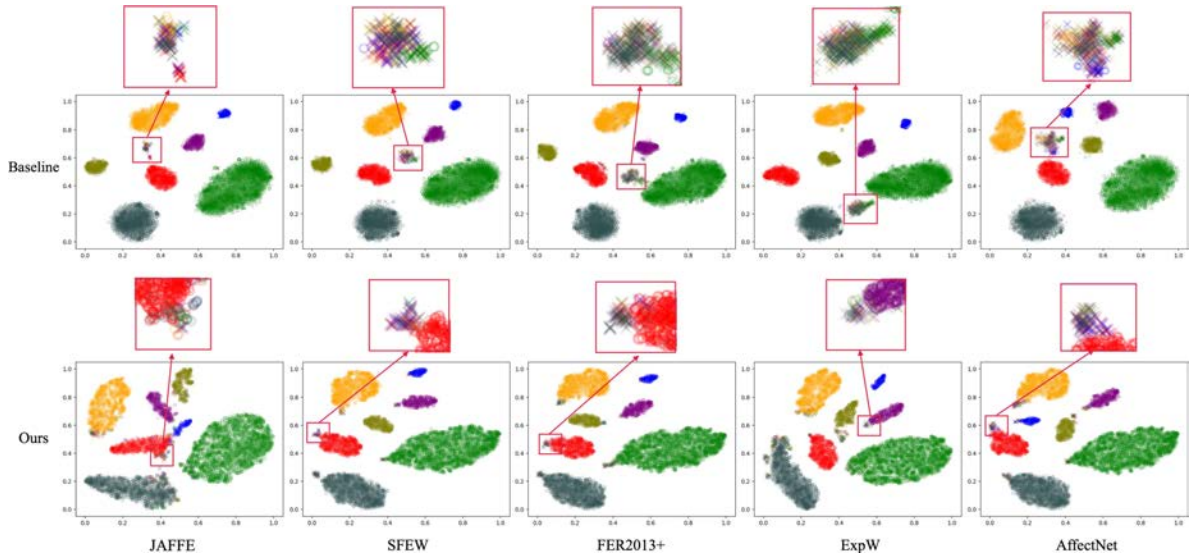


Fig. 11. Visualization of the features learned by the baseline and our approach. '$Circle$' and '$\times$' are specific to source and target domains, respectively. Each color represents an expression category. RAFDB is used for training.

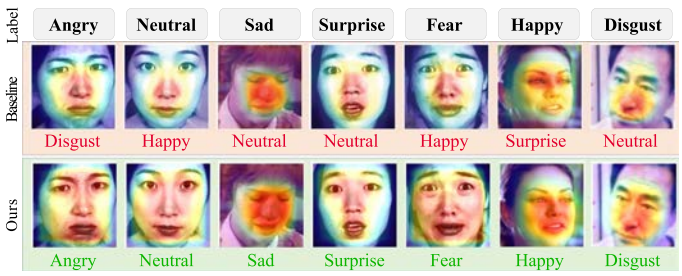


Fig. 12. The learned CAMs on facial expressions of seven categories. Warm and cool colors correspond to larger and lower attention values. The red and green label the wrongly and correctly classified categories, respectively. RAFDB is used for training, and each image randomly chosen from a target domain is used for the visualization. The upper row shows the ground truth labels.

attention to broader regions, covering more key expression-sensitive parts, which is helpful to reduce the overfitting on the source domain database.

V. Conclusion and Discussion

To address the problem of overfitting to the source domain in cross-database facial expression recognition (CDFER), we propose a generalization-enhanced paradigm for feature-wise and correlation attentions. Specifically, a multi-kernel competitive convolution module is developed in feature-wise attention to explore more sub-salient features and provide a dynamic weight specific to each kernel for comprehensively representing

unseen samples. Meanwhile, a Monte Carlo-based dropout is introduced in feature-correlation attention by taking into account the characteristics of feature maps in reducing their redundancy. As far as we know, this work is one of the pioneer works to specifically enhance the generalization capacity of the attention mechanism. Extensive experiments on six public FER databases, including the comparison with additional works [61], [62], [63], [64], [65] in the supplementary materials, demonstrate that our approach outperforms state-of-the-art methods in the scenarios of domain generalization and domain adaptation for CDFER. Ablation studies and visualization results also show the usefulness of each module.

In our future work, we will resort to the identification of outlier samples and reduction of long-tail distribution influence to further enhance this cross-database generalization capacity. Quantitative evaluation of cross-database performance in terms of the generalization metrics will also be explored.

Furthermore, our method is still limited by certain shortcomings: (i) Monte Carlo sampling assumes that the input data should obey a specific distribution; if the actual distribution of the data deviates from this assumption, the resulting mask from sampling may not be ideal. (ii) Utilizing Monte Carlo sampling for dropout can result in increased computational costs because it requires multiple forward passes to estimate uncertainty, thereby bring additional runtime overhead for model training and inference. (iii) Since our method is mainly developed for cross-database scenario, enhancing the generalization performances in both the single-database and cross-database scenarios, or trading off these performances, is worth exploring.

Acknowledgment

The authors would like to thank the anonymous reviewers for their helpful comments and constructive suggestions.

References

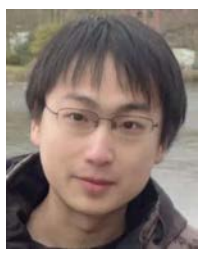
[1] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, "Domain generalization with mixstyle," in *Proc. Int. Conf. Learn. Representations*, 2021.
[2] W. Lu, J. Wang, H. Li, Y. Chen, and X. Xie, "Domain-invariant feature exploration for domain generalization," *Trans. Mach. Learn. Res.*, 2022.
[3] F. Qiao, L. Zhao, and X. Peng, "Learning to learn single domain generalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12556–12565.
[4] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2921–2929.
[5] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
[6] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021.
[7] Y. Ji, Y. Hu, Y. Yang, and H. T. Shen, "Region attention enhanced unsupervised cross-domain facial emotion recognition," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 4, pp. 4190–4201, Apr. 2023.
[8] M.-H. Guo et al., "Attention mechanisms in computer vision: A survey," *Comput. Vis. Media*, vol. 8, no. 3, pp. 331–368, 2022.
[9] W. Wu et al., "Boosting the transferability of adversarial samples via attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1161–1170.
[10] Y. Rao, G. Chen, J. Lu, and J. Zhou, "Counterfactual attention learning for fine-grained visual categorization and re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 1025–1034.
[11] X. Qu et al., "Attend to where and when: Cascaded attention network for facial expression recognition," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 6, no. 3, pp. 580–592, Jun. 2022.
[12] H. Liu, H. Wu, W. Xie, F. Liu, and L. Shen, "Group-wise inhibition based feature regularization for robust classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 478–486.
[13] S. Yang, W. Liu, Y. Yu, H. Hu, D. Chen, and T. Su, "Diverse feature learning network with attention suppression and part level background suppression for person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 1, pp. 283–297, Jan. 2023.
[14] Z. Zhao, Q. Liu, and S. Wang, "Learning deep global multi-scale and local attention features for facial expression recognition in the wild," *IEEE Trans. Image Process.*, vol. 30, pp. 6544–6556, 2021.
[15] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.
[16] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 510–519.
[17] H. Zhang, K. Zu, J. Lu, Y. Zou, and D. Meng, "EPSANet: An efficient pyramid squeeze attention block on convolutional neural network," in *Proc. Asian Conf. Comput. Vis.*, 2022, pp. 1161–1177.
[18] M.-H. Guo, Z.-N. Liu, T.-J. Mu, and S.-M. Hu, "Beyond self-attention: External attention using two linear layers for visual tasks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 5436–5447, May 2023.
[19] Y. Rao, W. Zhao, B. Liu, J. Lu, J. Zhou, and C.-J. Hsieh, "DynamicViT: Efficient vision transformers with dynamic token sparsification," in *Proc. 35th Int. Conf. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 13937–13949.
[20] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[21] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "DropBlock: A regularization method for convolutional networks," *Adv. Neural Inf. Process. Syst.*, vol. 31, pp. 10750–10760, 2018.
[22] S. Hou and Z. Wang, "Weighted channel dropout for regularization of deep convolutional neural network," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, pp. 8425–8432.
[23] B. Li et al., "Dropkey for vision transformer," in *Proc. 2023 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 22700–22709.
[24] M. Liu et al., "FocusedDropout for convolutional neural network," *Appl. Sci.*, vol. 12, no. 15, 2022, Art. no. 7682.
[25] Y. Zeng, T. Dai, B. Chen, S.-T. Xia, and J. Lu, "Correlation-based structural dropout for convolutional neural networks," *Pattern Recognit.*, vol. 120, 2021, Art. no. 108117.
[26] L. Li et al., "Progressive domain expansion network for single domain generalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 224–233.
[27] J. Zhang et al., "When neural networks fail to generalize? A model sensitivity perspective," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 11219–11227.
[28] K. Yan et al., "Cross-domain facial expression recognition based on transductive deep transfer learning," *IEEE Access*, vol. 7, pp. 108906–108915, 2019.
[29] 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*, vol. 333, pp. 231–239, 2019.
[30] F. Ma, B. Sun, and S. Li, "Facial expression recognition with visual transformers and attentional selective fusion," *IEEE Trans. Affect. Comput.*, vol. 14, no. 2, pp. 1236–1248, Apr.–Jun. 2023.
[31] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
[32] A. Vaswani et al., "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 6000–6010, 2017.
[33] F. Xue, Q. Wang, and G. Guo, "TransFER: Learning relation-aware facial expression representations with transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3601–3610.
[34] H. Zhang, D. Qu, K. Shao, and X. Yang, "DropDim: A regularization method for transformer networks," *IEEE Signal Process. Lett.*, vol. 29, pp. 474–478, 2022.
[35] H. Wang, G. Wang, G. Li, and L. Lin, "CamDrop: A new explanation of dropout and a guided regularization method for deep neural networks," in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manage.*, 2019, pp. 1141–1149.
[36] Y. Ding, S. Dong, Y. Tong, Z. Ma, B. Xiao, and H. Ling, "Channel DropBlock: An improved regularization method for fine-grained visual classification," in *Proc. 32nd Brit. Mach. Vis. Conf.*, 2021, pp. 1–13.
[37] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.
[39] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," *Adv. Neural Inf. Process. Syst.*, vol. 27, pp. 3320–3328, 2014.
[40] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 356–370, Jan. 2019.
[41] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *Proc. 18th ACM Int. Conf. Multimodal Interaction*, 2016, pp. 279–283.
[42] 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 *Proc. ACM Int. Conf. Multimodal Interaction*, 2015, pp. 423–426.
[43] 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.*, vol. 10, no. 1, pp. 18–31, Jan.–Mar. 2019.
[44] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "From facial expression recognition to interpersonal relation prediction," *Int. J. Comput. Vis.*, vol. 126, no. 5, pp. 550–569, 2018.
[45] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Proc. 3rd IEEE Int. Conf. Autom. Face Gesture Recognit.* IEEE, 1998, pp. 200–205.
[46] A. Dhall et al., "Collecting large, richly annotated facial-expression databases from movies," *IEEE Multimedia*, vol. 19, no. 3, pp. 34–41, Jul.–Sep. 2012.
[47] I. J. Goodfellow et al., "Challenges in representation learning: A report on three machine learning contests," in *Proc. Int. Conf. Neural Inf. Process.* Springer, 2013, pp. 117–124.
[48] A. Ballas and C. Diou, "Cnns with multi-level attention for domain generalization," in *Proc. 2023 ACM Int. Conf. Multimedia Retrieval*, 2023, pp. 592–596.
[49] Y. Tatsunami and M. Taki, "Sequencer: Deep LSTM for image classification," *Adv. Neural Inf. Process. Syst.*, pp. 38204–38217, 2022.
[50] C.-Y. Lee, T. Batra, M. H. Baig, and D. Ulbricht, "Sliced wasserstein discrepancy for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10285–10295.
[51] S. Li and W. Deng, "A deeper look at facial expression dataset bias," *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 881–893, Apr.–Jun. 2022.
[52] K. Fatras, T. Séjourné, R. Flamary, and N. Courty, "Unbalanced minibatch optimal transport; applications to domain adaptation," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 3186–3197.
[53] B. Peng, M. Islam, and M. Tu, "Angular gap: Reducing the uncertainty of image difficulty through model calibration," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 979–987.
[54] W. Zhu, J.-L. Yin, B.-H. Chen, and X. Liu, "SRoUDA: Meta self-training for robust unsupervised domain adaptation," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 3852–3860.
[55] D. Zhao, S. Liu, Y. Chen, W. Ji, and S. Ni, "Uncertainty learning facial expression recognition based on Monte-Carlo dropout," in *Proc. 2023 7th Asian Conf. Artif. Intell. Technol.*, 2023, pp. 1529–1535.
[56] J.-H. Nam and S.-C. Lee, "Random image frequency aggregation dropout in image classification for deep convolutional neural networks," *Comput. Vis. Image Understanding*, vol. 232, 2023, Art. no. 103684.
[57] B. Xie, L. Yuan, S. Li, C. H. Liu, X. Cheng, and G. Wang, "Active learning for domain adaptation: An energy-based approach," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, pp. 8708–8716.
[58] Y. Zhang et al., "Free lunch for domain adversarial training: Environment label smoothing," in *Proc. Int. Conf. Learn. Representations*, 2023.
[59] X. Shen, X. Tian, T. Liu, F. Xu, and D. Tao, "Continuous dropout," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 9, pp. 3926–3937, Sep. 2018.
[60] L. Van Der Maaten, "Accelerating t-SNE using tree-based algorithms," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3221–3245, 2014.
[61] Y. Zhang, C. Wang, X. Ling, and W. Deng, "Learn from all: Erasing attention consistency for noisy label facial expression recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 418–434.
[62] X. Jin, C. Lan, W. Zeng, and Z. Chen, "Style normalization and restitution for domain generalization and adaptation," *IEEE Trans. Multimedia*, vol. 24, pp. 3636–3651, 2022.
[63] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2584–2593.
[64] S. Roy and A. Etemad, "Active learning with contrastive pre-training for facial expression recognition," in *Proc. Int. Conf. Affect. Comput. Intell. Interaction*, 2023, pp. 1–8.
[65] H. Rangwani, S. K. Aithal, M. Mishra, A. Jain, and R. V. Babu, "A closer look at smoothness in domain adversarial training," in *Proc. 39th Int. Conf. Mach. Learn.*, 2022, pp. 18378–18399.

Weicheng Xie (Member, IEEE) received the B.S. degree in statistics from Central China Normal University, Wuhan, China, in 2008, the M.S. degree in probability and mathematical statistics, and the Ph.D. degree in computational mathematics from Wuhan University, Wuhan, in 2010 and 2013, respectively. He is currently an Associate Professor with the School of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. He is currently Visiting Research Fellow with School of Computer Science, University of Nottingham, Nottingham, U.K. His research interests include facial expression analysis and robust network design.

Tao Zhong received the B.Sc. degree in computer science from the Guangdong University of Technology, Guangzhou, China, in 2021. He is currently working toward the M.Sc. degree with the School of Computer Science and Technology, Shenzhen University, Shenzhen, China. His research focuses on cross-database expression analysis.

Fan Yang received the B.Sc. degree from the School of Computer Science and Technology, Tianjin Polytechnic University, Tianjin, China, in 2021. He is currently working toward the M.Sc. degree with the Computer Science and Technology, Shenzhen University. His research interests include facial expression recognition and action unit detection.



Siyang Song is currently a Lecturer (Assistant Professor) with the University of Exeter, Exeter, U.K. He is also an affiliated Researcher with the Department of Computer Science and Technology, University of Cambridge, Cambridge, U.K. His current research interests include affective computing, graph representation learning, computer vision, and machine learning.

Zitong Yu received the Ph.D. degree in computer science and engineering from the University of Oulu, Oulu, Finland, in 2022. He is currently an Assistant Professor with Great Bay University, China. He was a Postdoctoral Researcher with ROSE Lab, Nanyang Technological University, Singapore. From July to November 2021, he was a Visiting Scholar with TVG, University of Oxford, Oxford, U.K. His research interests include computer vision and biometric security. He was the recipient of the IAPR Best Student Paper Award, IEEE Finland Section Best Student Conference Paper Award 2020, Second Prize of the IEEE Finland Jt. Chapter SP/CAS Best Paper Award (2022), and the World's Top 2% Scientists 2023 by Stanford.

Linlin Shen (Senior Member, IEEE) is currently a Pengcheng Scholar Distinguished Professor with the School of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. He is a Honorary Professor with the School of Computer Science, University of Nottingham, Nottingham, U.K. He is the Deputy Director of National Engineering Lab of Big Data Computing Technology, Director of Computer Vision Institute, AI Research Center for Medical Image Analysis and Diagnosis and China-UK Joint Research Lab for Visual Information Processing. His research interests include deep learning, facial recognition, analysis/synthesis and medical image processing. Prof. Shen is listed as the "Most Cited Chinese Researchers" by Elsevier, "Top 0.05% Highly Ranked Scholar" by ScholarGPS, and listed in a ranking of the "Top 2% Scientists in the World" by Stanford University. He was the recipient of the "Best Paper Runner-up Award" from the Journal of IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, and "Most Cited Paper Award" from the Journal of Image and Vision Computing. His cell classification algorithms were the winners of the International Contest on Pattern Recognition Techniques for Indirect Immunofluorescence Images held by ICIP and ICPR. His team has also been the runner-up and second runner-up of a number of competitions for object detection in remote sensing images, nucleus detection in histopathology images, and facial expression recognition. He is the Co-Editor-in-Chief of the *IET Journal of Cognitive Computation and Systems* and the Senior Editor of *Expert Systems With Applications*.