# Learning Deep Global Multi-Scale and Local Attention Features for Facial Expression Recognition in the Wild

Zengqun Zhao, *Graduate Student Member, IEEE*, Qingshan Liu, *Senior Member, IEEE*, and Shanmin Wang

***Abstract*— Facial expression recognition (FER) in the wild received broad concerns in which occlusion and pose variation are two key issues. This paper proposed a global multi-scale and local attention network (MA-Net) for FER in the wild. Specifically, the proposed network consists of three main components: a feature pre-extractor, a multi-scale module, and a local attention module. The feature pre-extractor is utilized to pre-extract middle-level features, the multi-scale module to fuse features with different receptive fields, which reduces the susceptibility of deeper convolution towards occlusion and variant pose, while the local attention module can guide the network to focus on local salient features, which releases the interference of occlusion and non-frontal pose problems on FER in the wild. Extensive experiments demonstrate that the proposed MA-Net achieves the state-of-the-art results on several in-the-wild FER benchmarks: CAER-S, AffectNet-7, AffectNet-8, RAFDB, and SFEW with accuracies of 88.42%, 64.53%, 60.29%, 88.40%, and 59.40% respectively. The codes and training logs are publicly available at https://github.com/zengqunzhao/MA-Net.**

***Index Terms*— Facial expression recognition, deep convolutional neural networks, multi-scale, local attention.**

## I. Introduction

FACIAL expression, one of the most powerful and natural signals for human beings to convey their emotions, plays a significant role in communication. Automatic facial expression recognition (FER) has become an increasing fascinating topic in computer vision due to its applications in various fields, such as human-computer interaction (HCI) [1], driver fatigue monitoring [2], medical diagnosis [3], [4], and so on. FER aims to classify an image or a video clip into one of several basic emotions, i.e., neutral, happiness, sadness, surprise, fear, disgust, anger, and even contempt sometimes. FER can be divided into static FER and dynamic FER according to the input of the image or video, and laboratory-controlled and in-the-wild condition according to the scenario. In recent years, FER has achieved amazing performance on laboratory-controlled datasets [5]–[11], such as CK+ [12] and JAFER [13] facial expression datasets, which human faces are all frontal without any occlusion. However, problems of illumination variation, occlusion, and pose variation are challenging FER when the recognition scenario is transferred from laboratory-controlled to in-the-wild condition. Performance on in-the-wild datasets, such as CAER-S [14], RAF-DB [15], AffectNet [16], and SFEW [17] are greatly inferior to laboratory-controlled datasets [14], [18]–[21].

To improve the performance of FER in the wild, it is important and necessary to address the issues of occlusion and pose variation. Due to the fact that the occlusion and pose variation problems lead to a significant change of facial appearance in spatial level, the relevant studies focus mainly on image-based FER in the wild. The early work [22]–[25] made some efforts for addressing the occlusion problem by reconstructing the occluded geometric or textured features. Bourel *et al.* [22] proposed the enhanced Kanade-Lucas tracker to recover lost or drifted facial points. PCA-based methods were employed to reconstruct the positions of missing points in [23] and [24]. Hammal *et al.* [25] proposed a modified transferable belief model (TBM) to recognize facial expressions from partially occluded images. However, it is difficult to reconstruct occlusion regions in the real world well. With the prevalence of deep learning and the collection of in-the-wild datasets, much work employed deep convolutional neural networks (CNNs) to address the issues of occlusion as well as pose variation. For occlusion problems, patch-based methods are effective, which could capture the importance of facial patches, and how to select facial patches is a key issue to these methods. Some methods selected facial patches of interest relied on facial landmarks [18], [26], [27]. Wang *et al.* [20] employed three kinds of patch generation schemes, namely, fixed position cropping, random cropping, and landmark-based cropping, which could alleviate the problem of pose variation as well. Selecting local regions according to facial landmarks or randomly cropping may result in misalignment or uncertainty. For pose variation, some methods performed pose normalization before FER [28], [29]. Zhang *et al.* [30] proposed a method

Manuscript received January 21, 2021; revised May 11, 2021 and June 13, 2021; accepted June 23, 2021. Date of publication July 5, 2021; date of current version July 22, 2021. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61825601 and in part by the Natural Science Foundation of Jiangsu Province (NSF-JS) under Grant BK20192004B. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Soma Biswas. *(Corresponding author: Qingshan Liu.)*

Zengqun Zhao and Qingshan Liu are with the Engineering Research Center of Digital Forensics, Ministry of Education, Nanjing University of Information Science and Technology, Nanjing 210044, China (e-mail: zqzhao@nuist.edu.cn; qsliu@nuist.edu.cn).

Shanmin Wang is with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China, and also with the Engineering Research Center of Digital Forensics, Ministry of Education, Nanjing 210044, China (e-mail: smwang1994@163.com).

Digital Object Identifier 10.1109/TIP.2021.3093397

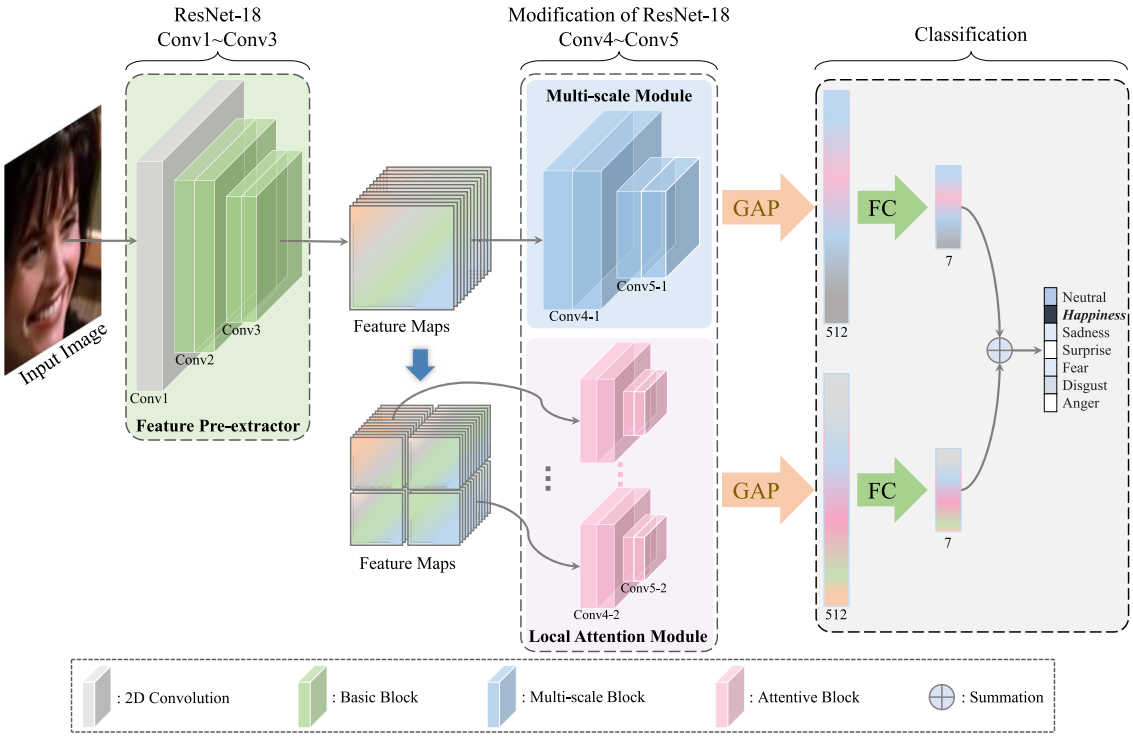



Fig. 1. The structure of the proposed method. The method consists of three components, including a feature pre-extractor, a multi-scale module, and a local attention module. GAP denotes global average polling, and FC denotes a fully-connected network.

to train a single FER classifier with multi-pose examples. Wang *et al.* [31] proposed an adversarial feature learning method to address pose variation as well as identity bias. However, the above methods require large amounts of training data with varying poses.

Learning facial features from various perspectives may achieve better performance under occlusion and pose variation conditions, and the study on psychology indicates that human face perception mechanisms extract both holistic and part information [32]. To this end, we propose a global multi-scale and local attention network (MA-Net) from the global and local perspective to acquire robust features, which can address both occlusion and pose variation problems. In CNNs, deeper convolution has a wider receptive field with rich semantic features, while shallower convolution has a narrow receptive field with rich geometry features [33]. The wider receptive field in deeper convolution is susceptible to occlusion and variant pose while adding shallower geometry features can decrease the susceptibility effectively so that the network can learn more comprehensive features. Therefore, the multi-scale module is designed to extract features with different receptive fields, which increases the diversity and robustness of global features. Instead of acquiring multi-scale features in a layer-wise manner, inspired by Res2Net [34], we extract multi-scale features within a single basic block. Also, the features extracted from local salient regions are critical to solve the issues of occlusion and non-frontal pose. Therefore, the local attention module is designed to extract local salient features, which releases the interference of occlusion and non-frontal pose situations. It is worth mentioning that previous works in extraction of local features by selecting local regions according to facial landmarks [18] or cropping [20] may result in misalignment or uncertainty. Different from these methods, ours is to divide the pre-extracted feature maps into several local regions without overlap directly, comparatively simple but efficient. Fig. 1 illustrates the main idea of the proposed method.

As shown in Fig. 1, we design a two-branch network for learning global multi-scale and local attention features. In the first branch, a multi-scale module is devised to learn robust multi-scale features towards occlusion and non-frontal pose conditions. In the second branch, the extracted feature maps are first divided into several local feature maps along the spatial axis without overlap, and then these local feature maps are processed by several parallel local attention networks. Finally, we employ a decision-level fusion to obtain recognition results. The contributions of our work can be summarized as follows:

1) We propose a global multi-scale and local attention network (MA-Net) for facial expression recognition in the wild. The proposed MA-Net is capable of acquiring robust both global and local features, which can address the issues both of occlusion and pose variation well.
2) The multi-scale module is designed to extract multi-scale features within a single basic block, and it can decrease the susceptibility of deeper convolution towards occlusion and variant pose effectively.
3) The local attention module is proposed to focus on local salient features, and it can release the interference of occlusion and non-frontal pose problems.

4) Experiments conducted on realistic occlusion and pose variation test datasets demonstrate that the proposed MA-Net attains strong robustness under occlusion and pose variation conditions. Moreover, our MA-Net also achieves state-of-the-art results on several benchmarks, and codes are publicly available.

## II. Related Work

This paper aims at static FER in the wild, in which occlusion and pose are two key issues, so we mainly focus on image-based FER. Image-based FER methods can be categorized into two parts: holistic-based FER and region-based FER.

### A. Holistic-Based FER

Holistic-based methods treat the face as a whole and focus mainly on global features. Early work mainly utilized hand-crafted features or shallow learning, such as SIFT [35], HOG [36], Gabor wavelet coefficients [37], and sparse learning [38]. Instead of designing effective features, Yang *et al.* [39] proposed a novel RankBoost method for FER enhanced by estimated facial expression intensity. With the rapid progress of deep learning, many methods, proposed using deep neural networks, improved the accuracy of FER in two ways: enhancing the complexity of the FER model and designing novel loss functions.

To obtain excellent features by enhancing the complexity of the FER model, Ding *et al.* [40] proposed FaceNet2ExpNet, a two-stage training algorithm for FER. In the first stage, a probabilistic distribution function was proposed to model the high-level neuron response. In the second stage, they performed label supervision to boost the final discriminative capability. Xie *et al.* [6] proposed the DAM-CNN model utilizing a SERD module to adaptively highlight the features that were highly relevant to the FER task, and MPVS-Net based on the encoder-decoder architecture to handle different variations. Hayale *et al.* [41] proposed deep siamese neural networks with a supervised loss function to embed verification and identification signals into a facial expression recognition pipeline. The latest related work proposed wide ensemble-based convolutional neural networks named ESR-9 [42], which could reduce the redundancy and computational load dramatically compared with the ensemble deep networks.

To enhance the discriminative power of the learned features by designing novel loss functions, Cai *et al.* [43] proposed an island loss to reduce the intra-class variations while enlarging the inter-class differences simultaneously. Zeng *et al.* [44] proposed a new feature loss to embed the information of hand-crafted features into the training process of the network. The latest related work proposed a separate loss [19] to maximize the intra-class similarity while minimizing the similarity between different classes.

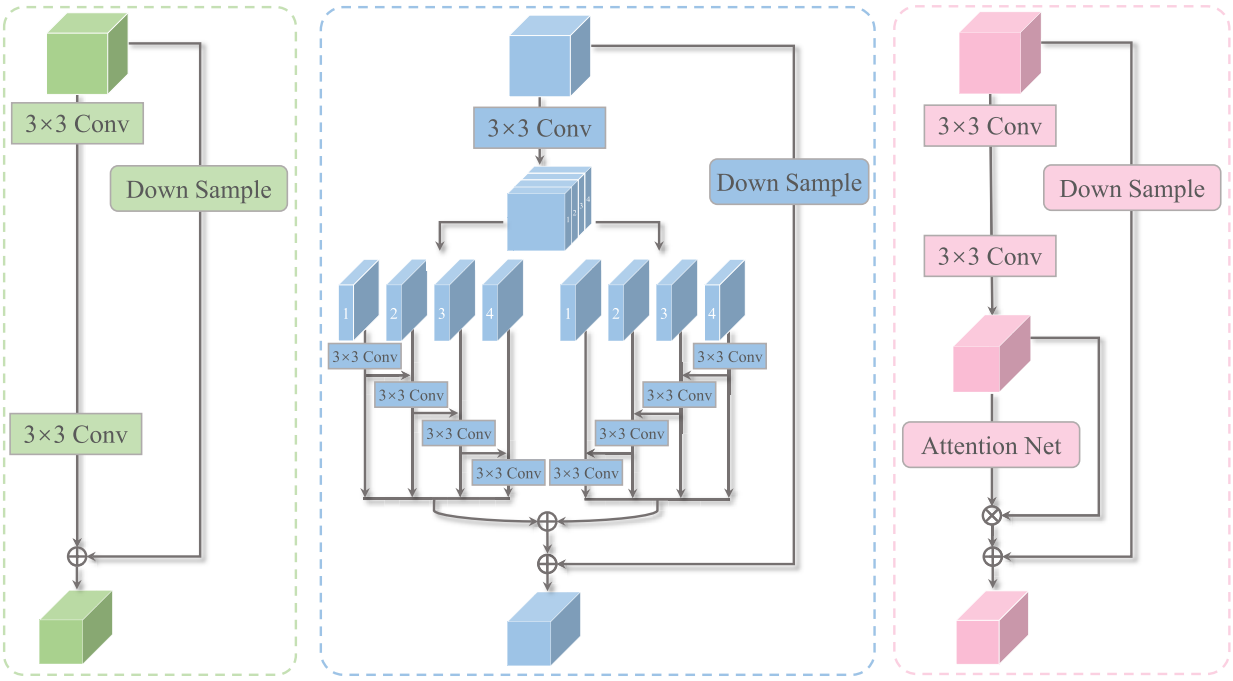
To address the problems of occlusion and pose variation, some researchers proposed occlusion- and pose-robust methods for FER from a global perspective. To obtain the occlusion-robust FER model, previous work mainly attempt to reconstruct the occluded geometric or texture features. Mao *et al.* [23] firstly detected facial occlusion based on the robust principal component analysis (RPCA) and then reconstructed occlusion and reweight AdaBoost classification. Jiang and Jia [24] also reconstructed occlusion using RPCA, and then utilized Eigenfaces and Fisherfaces to extract facial expression features, respectively. Hammal *et al.* [25] proposed a modified transferable belief model (TBM) to recognized facial expressions from partially occluded images. However, explicitly reconstruct occlusion regions in the real-world is complicated. With the popularity of deep learning, Pan *et al.* [45] utilized non-occluded facial images as privileged information, and two CNNs were trained from occluded and non-occluded facial images respectively. Then they fixed the non-occluded network on guiding the fine-tuning of the occluded network.

For the pose variation problem, some methods performed pose normalization before FER [28], [29]. Zhang *et al.* [30] utilized an encoder-decoder structure to synthesize facial images, which is expected to train a single FER classifier with multi-pose. Zhang *et al.* [21] promoted it by adding a facial geometry embedding network to extract the geometry vector. Wang *et al.* [31] proposed an adversarial feature learning method to address pose variation as well as identity bias. They employed a pose discriminator and a subject discriminator to classify the pose and the subject from the extracted feature representations respectively.

### B. Region-Based FER

Region-based methods divide the face into several overlapped or non-overlapped local regions and pay attention to local features. Zhong *et al.* [46] discovered some common and specific patches for facial expression. The common patches were used to recognize all expressions, while the specific patches were used for only a particular expression. Happy and Routray [26] selected 19 patches around eyes, nose, and mouth for facial expression recognition, and LBP features were extracted for each patch to train Support Vector Machines (SVMs). Liu *et al.* [47] extracted robust deep salient features from saliency-guided facial patches and fed features into a novel conditional CNNs enhanced random forest (CoNERF) to enhance decision trees. Li *et al.* [27] extracted 24 regions of interest according to face landmarks. Each patch is then processed by a Patch-Gated Unit (PG-Unit). A graph-structured representation was proposed by Zhong *et al.* [48]. Each node on the graph represents appearance information around the facial landmarks and edges represent the geometric information encoded by the distance between two nodes.

Some occlusion-robust methods also proposed for FER from a local perspective. Li *et al.* [27] proposed Patch-Gated CNN, which firstly decomposed facial images into several patches according to facial landmarks, and a patch gated unit with an attention net was designed to extract the features of each patch. Li *et al.* [18] promoted it by adding the whole face area to patch gated CNN. Wang *et al.* [20] selected some facial regions, both from facial landmarks and random cropping. Then, a self-attention module was employed to process each




(a) Basic Block (b) Multi-scale Block (c) Attentive Block

Fig. 2. Three types of the block are employed in MA-Net. The basic block, multi-scale block, and attentive block are utilized in the feature pre-extractor, the multi-scale module, and the local attention module, respectively.

facial region, and a relation-attention was employed to learn the weights of individual features.

## III. Methods

### A. Overview

As shown in Fig. 1, the proposed MA-Net consists of three components, including the feature pre-extractor, the multi-scale module, and the local attention module. The feature pre-extractor is to acquire middle-level facial features, and the feature pre-extractor consists of one 2D convolution layer and four basic blocks. The basic block structure shown in Fig. 2(a) is a basic building block utilized in ResNet-18 and ResNet-34 [49]. Then, a two-branch network is designed to process extracted feature maps, so that both global and local features can be obtained. In the first branch, we utilize a multi-scale module to learn global multi-scale features, which takes whole extracted feature maps as input. In the second branch, we first divide the extracted feature maps into several regional feature maps along the spatial axis without overlap, then, several parallel local attention networks are utilized to learn local salient features. The extracted multi-scale feature maps and local attention feature maps are followed by a global average pooling layer and a fully-connected network respectively. Finally, a decision-level fusion is utilized to obtain recognition results.

### B. Multi-Scale Module

Multi-scale feature representations of CNNs are critical to many vision tasks including object detection [50], [51], face analysis [52]–[55], semantic segmentation [56], [57], and so on. Most of these methods represent the multi-scale features in a layer-wise manner. Inspired by Res2Net [34], we design a multi-scale block that can acquire multi-scale features at a granular level within a single basic block.

Fig. 2(b) shows the structure of the proposed multi-scale block. As shown in Fig. 2(b), we introduce a symmetrical structure to learn multi-scale features within a basic block, such a method can ensure that the feature subset at the front and back both can contain richer scale information. Specifically, after the $3 \times 3$ convolution, the feature maps $X$ can be obtained. We evenly split the feature maps $X$ along the channel axis into $n$ feature map subsets denoted by $X_i$, where $i \in \{1, 2, \ldots, n\}$. Therefore, each feature map subset $X_i$ has the same spatial size but $1/n$ channels compared with the feature maps $X$. Then, each $X_i$ is processed by a corresponding $3 \times 3$ convolution denoted by $P_i^p(\cdot)$, where $p \in \{left, right\}$ denotes the position. $Y_i^p$ denotes the output of $P_i^p(\cdot)$. Hence, each output $Y_i^p$ can be written as:

$$Y_i^{left} = \begin{cases} P_i^{left}(X_i) & i = 1 \\ P_i^{left}(X_i + Y_{i-1}^{left}) & 1 < i \leqslant n \end{cases} \tag{1}$$

$$Y_i^{right} = \begin{cases} P_i^{right}(X_i) & i = n \\ P_i^{right}(X_i + Y_{i+1}^{right}) & 1 \leqslant i < n \end{cases} \tag{2}$$

Then, the final output $Y_i$ can be written as:

$$Y_i = Y_i^{left} + Y_i^{right} \tag{3}$$

From Eq. (1), we can notice that each operation of $P_i^{left}(\cdot)$ can capture features from all subsets $\{X_j, j \leqslant i\}$. And from Eq. (2) we can notice that each operation of $P_i^{right}(\cdot)$ can


Fig. 3. The comparison of class activation mapping (CAM) between the multi-scale module and traditional ResNet. The images are from the test set of the FED-RO and Pose-AffectNet datasets.

capture features from all subsets $\{X_j, n \geqslant j \geqslant i\}$. During one operation, a split feature $X_i$ is processed by a 3×3 convolution. The output $Y_i^{left}$ have a large receptive field than $\{Y_k, k < i\}$ and the output $Y_i^{right}$ have a large receptive field than $\{Y_k, k > i\}$. Hence, each output $Y_i^p$ contains subset features with a different number and different scale. To obtain more diverse multi-scale features, we concatenate all $Y_i^p$s along the channel axis. The larger $n$ potentially allows features to contain richer scale information, but it may boost computational overheads. In our work, we set $n = 4$, which makes a trade-off between performance and computation.

The multi-scale module consists of four multi-scale blocks and followed by a global average pooling (GAP) layer. After GAP, we can obtain a feature vector with a size of 512. To better explain the effect of the multi-scale module, we conduct visualization of the proposed module through class activation mapping (CAM) [58] to compare the performance of the multi-scale module with the traditional ResNet. As shown in Fig. 3, the stronger CAM areas are covered with the lighter colors.

Due to the multi-scale convolutions considering both deeper semantic and shallower geometry features, the learned multi-scale features not only enhances the diversity of global features but also reduces the susceptibility of the deeper convolutions towards occlusion and variant pose. Hence, the networks can obtain a more comprehensive representation of global features. Compared with the traditional ResNet, the multi-scale-based CAM results can pay attention to specific regions which are beneficial to facial expression recognition, even if there are occlusion and non-frontal pose issues in the facial images.

### C. Local Attention Module

For occlusion and pose variation conditions of FER in the wild, it is crucial to pay attention to local features. To acquire effective local features, previous methods mainly divide the face into several patches by facial landmarks or random cropping. These methods may result in misalignment or uncertainty to FER in the wild. While in our method, the mid-level feature maps are divided into several local feature maps without overlap, and each network for local feature maps is expected that can focus on local salient features autonomously by attention mechanism. Therefore, after the pre-extraction module, we divide the extracted feature maps $S$ into several local feature maps $S_i$ along the spatial axis, where $i \in \{1, 2, \ldots, m\}$. We consider that dividing the feature maps into four local feature maps conforms to the facial region related to expression. Hence, we set $m = 4$, and an ablative study in Sec. IV. C. also demonstrates that using four feature patches achieves the best accuracy. As a result, each local feature map $S_i$ has the same channels but $1/2$ spatial size compared with the original feature map $S$.

Specifically, after the pre-extractor module, the feature map is $S \in \mathbb{R}^{28\times28\times128}$, where 28 is the spatial size, and 128 is the channel size. Then, the extracted feature map is divided into 4 regional feature maps $S_i \in \mathbb{R}^{14\times14\times128}$ without overlap, where $i \in \{1, 2, 3, 4\}$.

Fig. 2(c) shows the structure of the proposed attentive block. After two 3 × 3 convolution, we can obtain feature maps denoted by $F \in \mathbb{R}^{H\times W\times C}$. Then, a convolutional block attention module (CBAM) [59] is employed as our attention net. The CBAM can sequentially infers attention maps along two separate dimensions, channel and spatial, then the attention maps are multiplied to the input feature map for adaptive feature refinement. Moreover, the CBAM is a lightweight and general module, it can be integrated into any CNN architectures seamlessly with negligible overheads.

In our network, the attention net takes $F$ as input and infers a 1D channel attention map $M_c \in \mathbb{R}^{1\times1\times C}$ and a 2D spatial attention map $M_s \in \mathbb{R}^{H\times W\times1}$. Therefore, the attention net can be formulated as:

$$F_r = M_s(F) \otimes (M_c(F) \otimes F) \tag{4}$$

where $\otimes$ denotes element-wise multiplication. Because the attention map and feature map have different dimension sizes, the channel attention values are broadcasted along the spatial dimension, and the spatial attention values are broadcasted along the channel dimension when conduct multiplication. $F_r$ is the final refined output.

The local attention module consists of four parallel local attention networks and each of which consists of four attentive blocks. The local attention module takes four local 14 × 14 × 128 feature maps as the input, and each local 14 × 14 × 128 feature maps are fed into a corresponding local attention network. With the local attention module, we can obtain four


Fig. 4. The class activation mapping (CAM) of the traditional ResNet, local feature module (the local feature module can be created by removing the attention net from the local attention module), and local attention module. The images are from the test set of the FED-RO and Pose-AffectNet datasets. LF denotes the local feature module, and LA denotes the local attention module.

local $7 \times 7 \times 512$ feature maps. The four local feature maps are then concatenated along the spatial axis, and the GAP layer is applied on the concatenated $14 \times 14 \times 512$ feature maps to obtain a feature vector with a size of 512.

Analogously, to better explain the effect of the local attention module, we also conduct visualization of CAM to validate the performance of the local attention module as well as the attention mechanism. As shown in Fig. 4, the images in the second and third rows are visualization consequences of local feature module and local attention module respectively. Compared with the traditional ResNet, the local-attention-based CAM results can guide the network to focus on the local salient regions which are crucial to enhance the robustness towards occlusion and non-frontal facial expression conditions. For example, the first four images are occluded faces, and the module can only focus on the non-occlusion regions, which is consistent with human perception. The last four images are non-frontal faces, and the local attention module is capable of focusing on the local salient regions. Compared with the local feature module, the attention-based method can enhance the significance of the local features and concentrate on the action units.

### D. Fusion Strategy and Loss Function

Feature-level fusion and decision-level fusion are two conventional methods [60]. The former directly combines the feature vectors of two branches into a joint feature vector and trains a classifier for FER [61]; while the latter combines recognition results from two branches. In our MA-Net, the extracted multi-scale features and local attention features have weak complementarity in the feature level. Therefore, the decision-level fusion strategy is employed in our research.

After the two GAP layers, we can obtain two feature vectors with a size of 512 which are donated by $v^{(k)}$, where $k \in \{local, global\}$ denotes branch. The loss function in our method is consist of two cross-entropy loss which can be formulated as:

$$\mathcal{L}_k = -\frac{1}{N}\sum_{i=0}^{N-1}\log\frac{e^{W_{y_i}^{(k)T}v_i^{(k)}+b_{y_i}^{(k)}}}{\sum_{j=0}^{C-1}e^{W_j^{(k)T}v_i^{(k)}+b_j^{(k)}}} \tag{5}$$

where N is the number of samples; C is the number of expression categories; $W^{(k)}$ is the weight matrix of the FC layer; $b^{(k)}$ is the bias term of the FC layer; $v_i^{(k)}$ is the FC input of the $i$th sample, and $y_i$ is its class label.

Then, the final loss function is defined as:

$$\mathcal{L} = \lambda\mathcal{L}_{local} + (1-\lambda)\mathcal{L}_{global} \tag{6}$$

where $\lambda$ is a hyper-parameters to balance two parts.

## IV. Experiments

### A. Datasets

We conduct the experiments on four popular in-the-wild facial expression datasets: CAER-S [14], RAF-DB [15], AffectNet [16], and SFEW [17], and five realistic occlusion and pose variation test sets: FED-RO [18], Occlusion-AffectNet, Occlusion-RAF-DB, Pose-AffectNet and Pose-RAF-DB.

*1) CAER-S:* The CAER-S [14] dataset was created by selecting static frames from the CAER dataset with 65,983 images and has been divided into two sets: training set (44,996 samples) and test set (20,987 samples). Each image is assigned to one of seven expressions, i.e., neutral, happiness, sadness, surprise, fear, disgust, and anger.

*2) RAF-DB:* The RAF-DB [15] dataset contains 30,000 facial images annotated with basic or compound expressions by 40 trained human coders. In our experiment, only images with basic emotions are used, including 12,271 images as training data and 3,068 images as test data.

*3) AffectNet:* The AffectNet [16] dataset is the largest dataset so far, and it provides both categorical and Valence-Arousal annotations. The dataset contains more than one million images collected from the Internet by querying

expression-related keywords in three search engines, in which 450,000 images are manually annotated with 11 expression categories. In our experiments, we recognize both seven and eight expression categories. The seven expression categories contain six basic expressions and neutral faces, while the eight expression categories with the addition of contempt expression. We select 283,901 images as training data and 3,500 images as test data when recognizing seven expression categories on AffectNet. The experimental setting is the same as IPA2LT [62], IPFR [31], and gACNN [18]. While recognizing eight expression categories, we select 287,568 images as training data and 4,000 images as test data. The experimental setting is the same as RAN [20], ESR-9 [42], and SCN [63].

*4) SFEW:* The SFEW [17] dataset is created by selecting static frames from the AFEW dataset after computing key-frames based on facial point clustering. The dataset covers unconstrained facial expressions, varied head poses, large age range, occlusion, varied focus, different resolution of the face, and real-world illumination. The most commonly used version is SFEW 2.0. It has been divided into three sets: training set (958 samples), validation set (436 samples), and the test set (372 samples). Each image is assigned to one of seven expression categories as CAER-S. Due to the annotations of the test set are not released, we mainly report our performance on the validation set.

*5) FED-RO:* To address the occlusion problem, Li *et al.* [18] collected and annotated a facial expression dataset with real occlusion (FED-RO) in the wild. They collected this dataset by mining Bing & Google search engine for occluded images. Each image was carefully labelled by three people. To ensure the images in FEO-RO are not included in RAF-DB or AffectNet dataset, they filtered out repeated facial images. FED-RO contains 400 images in total, and the images were categorized into seven basic expressions.

*6) Occlusion-AffectNet & Occlusion-RAF-DB:* To examine the performance of the FER model under real-world occlusion condition, Wang *et al.* [20] built two subsets, Occlusion-AffectNet and Occlusion-RAF-DB, from the validation set of AffectNet and the test set of RAF-DB respectively. These test sets are annotated with different occlusion types, and the Occlusion-AffectNet and Occlusion-RAF-DB contain 683 and 735 images in total respectively.

*7) Pose-AffectNet & Pose-RAF-DB:* To examine the performance of the FER model under variant pose condition, Wang *et al.* [20] also built two subsets, Pose-AffectNet and Pose-RAF-DB, from the validation set of AffectNet and the test set of RAF-DB respectively. The pitch or yaw angle of faces on Pose-AffectNet and Pose-RAF-DB are all larger than 30°. The Pose-AffectNet contains 1,948 and 985 faces with an angle larger than 30° and 45° respectively in total, and the Pose-RAF-DB contains 1,248 and 558 faces with an angle larger than 30° and 45° respectively in total.

### B. Implementation Details

For all the datasets, face images are detected and aligned using Retinaface [64] and then cropped and resized to 224 × 224 pixels. Random cropping and random horizontal flipping are employed to avoid over-fitting. The ResNet-18 [49] is employed as a backbone CNN. Our method is implemented with Pytorch toolbox [65] on one GeForce RTX 2080 Ti platform, using the SGD optimizer with a momentum of 0.9. The number of Parameters and FLOPs of our MA-Net is 50.54 M and 3.65 G, respectively.

For the RAF-DB and the AffectNet datasets, consistent with RAN [20] and SCN [63], we pre-train our MA-Net on MS-Celeb-1M [66] face recognition dataset and then fine-tune it on RAF-DB and AffectNet with spending of 1.0 and 4.5 hours approximately. For the SFEW dataset, consistent with Island Loss [43] and IPFR [31], we pre-train our MA-Net on the FER2013 [67] dataset and then fine-tune it on SFEW with a spending of 40 minutes approximately. For the CAER-S dataset proposed recently, we train our MA-Net from scratch with a batch size of 128, initializing the learning rate as 0.1 and dividing it by two every 50 epochs. The training operation is stopped in the 400th epoch and spends 14.0 hours approximately. For the Occlusion-AffectNet, the Occlusion-RAF-DB, the Pose-AffectNet, and the Pose-RAF-DB test set, we train our MA-Net using the same setting as RAN [20]. In the inference phase, our MA-Net achieves a recognition time of 0.0489 s for a single facial image, which runs at 20.45 FPS.

### C. Ablation Analysis

To validate the effectiveness of each component in our MA-Net, we conduct an ablation analysis on two benchmarks (CAER-S and RAF-DB) and two occlusion and pose variation test sets (FED-RO and Pose-AffectNet). In our experiments, the multi-scale module, local attention module, fusion strategy, and value of weight $\lambda$ are studied respectively, in which the two modules are implemented both in the one-branch and two-branch networks.

*1) Multi-Scale Module:* We first conduct the experiments to validate the effectiveness of the multi-scale module. Specifically, for the one-branch network, we replace the basic block of the ResNet-18 with the multi-scale block of the Res2Net and ours symmetrical multi-scale block respectively, and for the two-branch network, we replace one of the branches of the two-branch baseline with the proposed multi-scale module. The one-branch baseline network is the ResNet-18, and the two-branch baseline network is the modified ResNet-18, in which each branch consists of the last two convolution stages of ResNet-18.

The results of the proposed symmetrical multi-scale block in one-branch and two-branch networks are shown in Tab. I and Tab. II, respectively. The results indicate that the performance of the symmetrical multi-scale block is better than the multi-scale block of the Res2Net. The proposed symmetrical multi-scale module used in the one-branch network improves the recognition rate by 1.60%, 1.15%, 2.75%, 2.26%, and 2.54% on CAER-S, RAF-DB, FED-RO, Pose-AffectNet (30°), and Pose-AffectNet (45°), respectively. On the two-branch network, the symmetrical multi-scale module improves the recognition rate by 0.45%, 1.08%, 2.25%, 0.56%, and 0.61% on CAER-S, RAF-DB, FED-RO, Pose-AffectNet (30°), and

TABLE I

EVALUATION OF EACH COMPONENT IN ONE-BRANCH NETWORK ON CAER-S, RAF-DB, FED-RO, AND POSE-AFFECTNET **WITHOUT PRE-TRAINING**

| Methods | CAER-S | RAF-DB | FED-RO | Pose-AffectNet ($\geqslant 30°$) | Pose-AffectNet ($\geqslant 45°$) |
|---|---|---|---|---|---|
| Baseline (One Branch) | 84.41 | 82.59 | 60.00 | 53.55 | 52.39 |
| Baseline + Multi-scale | 85.50 | 82.89 | 62.00 | 54.22 | 53.61 |
| Baseline + Symmetrical Multi-scale | 86.01 | 83.74 | 62.75 | 55.81 | 54.93 |
| Baseline + Local Feature | 86.12 | 83.87 | 62.00 | 54.68 | 54.02 |
| Baseline + Local Feature + Attention | 86.32 | 84.65 | 63.25 | 55.30 | 54.73 |

TABLE II

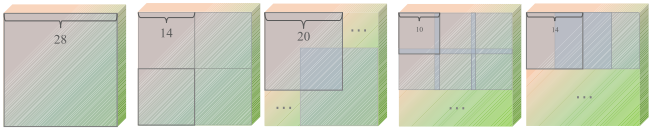
EVALUATION OF EACH COMPONENT IN OUR MA-NET ON CAER-S, RAF-DB, FED-RO, AND POSE-AFFECTNET **WITHOUT PRE-TRAINING**. THE LTS, MS, AND LA DENOTE THE LAST TWO CONVOLUTION STAGES OF RESNET-18, THE MULTI-SCALE MODULE, AND THE LOCAL ATTENTION MODULE, RESPECTIVELY

| Datasets | LTS | LTS | MS | LA | Acc.(%) |
|---|---|---|---|---|---|
| CAER-S | √ | √ | | | 87.23 |
| | √ | | √ | | 87.68 |
| | √ | | | √ | 87.98 |
| | | | √ | √ | **88.42** |
| RAF-DB | √ | √ | | | 83.05 |
| | √ | | √ | | 84.13 |
| | √ | | | √ | 85.43 |
| | | | √ | √ | **86.34** |
| FED-RO | √ | √ | | | 61.50 |
| | √ | | √ | | 63.75 |
| | √ | | | √ | 62.05 |
| | | | √ | √ | **65.00** |
| Pose-AffectNet ($\geqslant 30°$) | √ | √ | | | 53.86 |
| | √ | | √ | | 54.42 |
| | √ | | | √ | 56.17 |
| | | | √ | √ | **56.48** |
| Pose-AffectNet ($\geqslant 45°$) | √ | √ | | | 52.90 |
| | √ | | √ | | 53.51 |
| | √ | | | √ | 54.83 |
| | | | √ | √ | **55.95** |

Pose-AffectNet (45°), respectively. The results also show that the multi-scale module can boost the FER performance when the other branch employs the local attention module.

*2) Local Attention Module:* We then evaluate the validity of the local attention module. As shown in Tab. I, for the one-branch network, the local attention module improves the accuracy by 1.91%, 2.06%, 3.25%, 1.75%, and 2.34% on CAER-S, RAF-DB, FED-RO, Pose-AffectNet (30°), and Pose-AffectNet (45°), respectively. When the attention net is utilized in the one-branch network, the FER accuracies show an improvement of 0.20%, 0.78%, 1.25%, 0.62%, and 0.71% on CAER-S, RAF-DB, FED-RO, Pose-AffectNet (30°), and Pose-AffectNet (45°), respectively. The results shown in Tab. II indicate that, when the local attention module is utilized in the two-branch network, the FER accuracies show an improvement of 0.75%, 2.38%, 0.55%, 2.31%, and 1.93% on CAER-S, RAF-DB, FED-RO, Pose-AffectNet (30°), and Pose-AffectNet (45°), respectively.

To explore the impact of the local feature selection strategy on the local attention module, we evaluate three types of local feature selection strategy with different overlap presented in Fig. 5. Corresponding FER accuracy on the RAF-DB dataset shown in Tab. III. From the results, we can find that the accuracy of strategy without overlap is better than those with overlap. Furthermore, the best accuracy of the $3 \times 3$ strategy is 85.41%, which is inferior to the $2 \times 2$ strategy with the best result of 86.32%. We believe that too small regions lead to insufficient discrimination ability of local features. Dividing the feature maps into four local feature maps conforms to the facial region related to expression. These facial regions mainly include eyes, eyebrow corners, and lip corners.




(a) $1 \times 1$ (b) $2 \times 2$ (c) $3 \times 3$

Fig. 5. Three types of local feature selection strategy. The strategy of sub-figure (a) means that takes whole feature maps as an input, which can be denoted by $1 \times 1$. The strategy of sub-figure (b) means that divide the whole feature maps into 4 local regions without overlap or with an overlap of 6, which can be denoted by $2 \times 2$. The strategy of sub-figure (c) means that divide the whole feature maps into 9 local regions with an overlap of 1 or 7, which can be denoted by $3 \times 3$.

TABLE III

EVALUATION OF DIFFERENT REGION SELECTION STRATEGIES ON THE RAF-DB DATASET **WITHOUT PRE-TRAINING**

| Selection Strategies | # Regions | Acc.(%) |
|---|---|---|
| $1 \times 1$ (overlap = 0) | 1 | 84.58 |
| $2 \times 2$ (overlap = 0) | 4 | **86.32** |
| $2 \times 2$ (overlap = 6) | 4 | 85.14 |
| $3 \times 3$ (overlap = 1) | 9 | 85.41 |
| $3 \times 3$ (overlap = 7) | 9 | 85.29 |

TABLE IV

EVALUATION OF DIFFERENT FUSION STRATEGIES ON RAF-DB DATASET **WITHOUT PRE-TRAINING**

| Fusion Strategies | Acc.(%) |
|---|---|
| Feature-level | 82.76 |
| Decision-level | **86.34** |

*3) Fusion Strategy:* We also conduct experiment to evaluate different fusion strategies. As shown in Tab. IV, we compare two conventional fusion strategies: feature-level and decision-level. The feature-level fusion directly combines the feature vectors obtained by the two branches into a joint feature vector and trains a classifier for FER. While the decision-level fusion

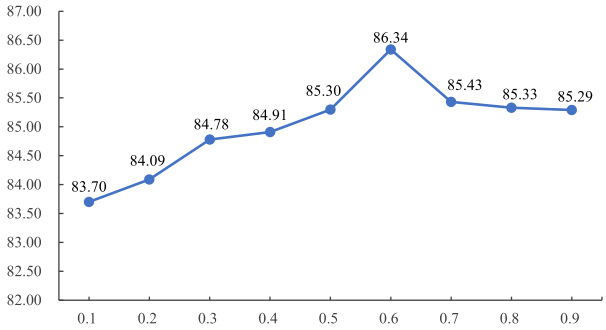



Fig. 6. Evaluation of different $\lambda$ values on RAF-DB dataset.

TABLE V

COMPARISON TO THE STATE-OF-THE-ART RESULTS ON THE CAER-S DATASET. * THE RESULT IS TRAINED FROM SCRATCH

| Methods | Years | Acc.(%) |
|---|---|---|
| ResNet-18 [49] | 2016 | 84.67 |
| ResNet-50 [49] | 2016 | 84.81 |
| Res2Net-50 [34] | 2019 | 85.35 |
| CAER-Net-S [14] | 2019 | 73.51 |
| MA-Net* (Ours) | 2020 | **88.42** |

combines recognition results from two branches. The results in Tab. IV indicate that the decision-level fusion is highly superior to the feature-level fusion strategies in our MA-Net. Intuitive cognition is that the extracted multi-scale features and local attention features have weak complementarity in the feature level.

*4) Weight $\lambda$:* $\lambda$ is a hyper-parameters to balance two parts of the loss function. To explore the impact of weight $\lambda$ for our MA-Net, we study different values from 0.1 to 0.9. Fig. 6 shows the results on the RAF-DB datasets. The results show that the best performance can be obtained when $\lambda = 0.6$, namely, the importance of the local attention branch is slightly higher than the multi-scale branch. Such a result also keeps in line with the ablation study result in Tab. I, in which the performance of the baseline with the local attention module is slightly better than the baseline with the multi-scale module.

### D. Comparison With State-of-the-Art Methods

In this section, we compare our best results to several state-of-the-art methods on CAER-S, AffectNet, RAF-DB and SFEW datasets.

*1) Comparison on CAER-S:* The FER accuracy on CAER-S achieved by MA-Net and its comparison with some state-of-the-art methods are shown in Tab. V. Due to the CAER-S dataset was proposed recently, and only [14] evaluated their method on it, we conduct several experiments utilizing some state-of-the-art networks on it, such as ResNet-18, ResNet-50, and Res2Net-50. The results in Tab. V demonstrate that our MA-Net outperforms state-of-the-art Methods, even though comparing with deeper networks, such as ResNet-50 and Res2Net-50. It is worth pointing out that our MA-Net is trained from scratch while others are pre-trained on ImageNet.

TABLE VI

COMPARISON TO THE STATE-OF-THE-ART RESULTS ON THE AFFECTNET DATASET. + OVERSAMPLING IS USED SINCE AFFECTNET IS IMBALANCED

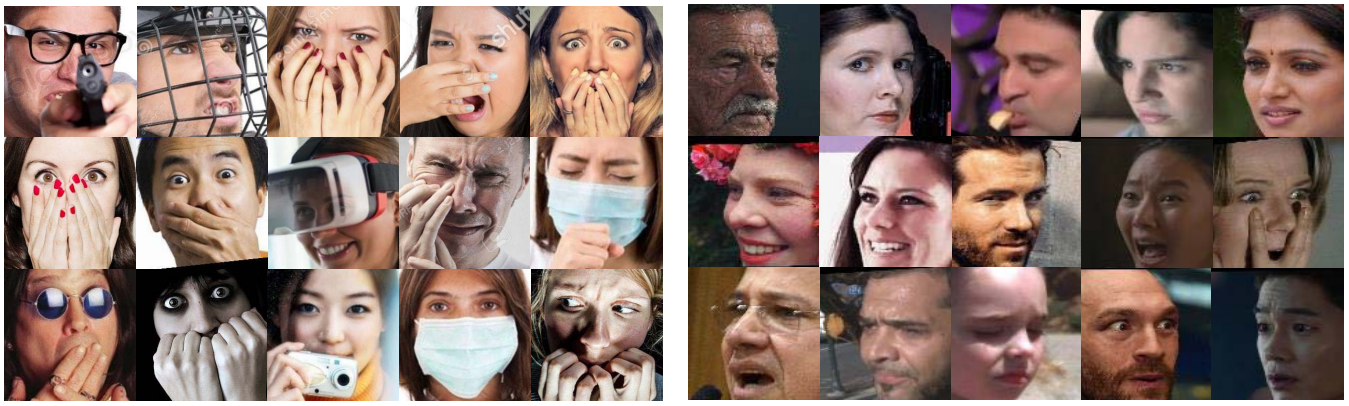
| Methods | Backbone | Years | # Classes | Acc.(%) |
|---|---|---|---|---|
| IPA2LT [62] | ResNet-80 | 2018 | 7 | 57.31 |
| gACNN [18] | VGG-16 | 2019 | 7 | 58.78 |
| IPFR [31] | Manually-Designed | 2019 | 7 | 57.40 |
| Separate Loss [19] | ResNet-18 | 2019 | 7 | 58.89 |
| FMPN [8] | Inception-V3 | 2019 | 7 | 61.52 |
| VGG-FACE [68] | VGG-Face | 2020 | 7 | 60.00 |
| SNA-DFER [69] | Manually-Designed | 2020 | 7 | 62.70 |
| LDL-ALSG [70] | ResNet-50 | 2020 | 7 | 59.35 |
| MA-Net+ (Ours) | ResNet-18 | 2020 | 7 | **64.53** |
| MobileNet-Variant [71] | Manually-Designed | 2018 | 8 | 56.00 |
| VGGNet-Variant [71] | Manually-Designed | 2018 | 8 | 58.00 |
| Weighted-Loss [16] | AlexNet | 2019 | 8 | 58.00 |
| RAN+ [20] | ResNet-18 | 2020 | 8 | 59.50 |
| ESR-9 [42] | Manually-Designed | 2020 | 8 | 59.30 |
| SCN+ [63] | ResNet-18 | 2020 | 8 | 60.23 |
| MA-Net+ (Ours) | ResNet-18 | 2020 | 8 | **60.29** |

Fig. 7. Facial images with the annotation of contempt were selected from the validation set of the AffectNet dataset.

*2) Comparison on AffectNet:* Different from other datasets, the AffectNet dataset is manually annotated with 11 expression categories, previous work recognize both 7 expression categories and 8 expression categories. To prove the effectiveness of our MA-Net sufficiently, we conduct experiments and compare with other state-of-the-art methods by classifying both 7 expression categories and 8 expression categories. Due to the AffectNet dataset has an imbalanced training set but a balanced validation set, we employ an oversampling strategy,[1] which is consistent with RAN [20] and SCN [63]. As shown in Tab. VI, we obtain 64.53% in term of FER accuracy on AffectNet with 7 expression categories, which is greatly superior to the state-of-the-art methods. We also obtain the highest accuracy of 60.29% on AffectNet with 8 expression categories.

In addition, our MA-Net is slightly superior to SCN [63] which achieves the result of 60.23%, while our MA-Net is significantly superior to other state-of-the-art methods. Moreover, the results of our MA-Net display a large gap between 7 expression categories and 8 expression categories, and the 8 expression categories added expression of contempt based on 7 expression categories. To this end, we randomly select some images from the AffectNet dataset with an annotation of contempt shown in Fig. 7. The randomly selected images

[1] https://github.com/ufoym/imbalanced-dataset-sampler



(a) FED-RO (b) Pose-AffectNet

Fig. 8. Some examples of occlusion and pose variation test datasets. The images of sub-figure (a) are from FER-RO dataset, which exist severe occlusion, the images of sub-figure (b) are from the test set of AffectNet, which exist severe pose variation.

TABLE VII

COMPARISON TO THE STATE-OF-THE-ART RESULTS ON THE RAF-DB DATASET. * THE RESULT ARE TRAINED FROM SCRATCH. † RAF-DB AND AFFECTNET ARE JOINTLY USED FOR TRAINING

| Methods | Backbone | Years | Acc. (%) |
|---|---|---|---|
| DLP-CNN [15] | 8-layer DCNN | 2017 | 84.22 |
| IPA2LT [62] | ResNet-80 | 2018 | 86.77 |
| Separate Loss [19] | ResNet-18 | 2019 | 86.38 |
| gACNN [18] | VGG-16 | 2019 | 85.07 |
| RAN [20] | ResNet-18 | 2020 | 86.90 |
| LDL-ALSG [70] | ResNet-50 | 2020 | 85.53 |
| SCN* [63] | ResNet-18 | 2020 | 78.31 |
| SCN [63] | ResNet-18 | 2020 | 87.03 |
| SCN† [63] | ResNet-18 | 2020 | 88.14 |
| MA-Net* (Ours) | ResNet-18 | 2020 | 86.32 |
| MA-Net (Ours) | ResNet-18 | 2020 | **88.40** |

TABLE VIII

COMPARISON TO THE STATE-OF-THE-ART RESULTS ON THE SFEW DATASET

| Methods | Pre-trained Dataset | Years | Acc. (%) |
|---|---|---|---|
| ADML [72] | FER-2013 | 2017 | 54.20 |
| IACNN [73] | FER-2013 | 2017 | 54.30 |
| DCD [74] | FER-2013 | 2018 | 49.18 |
| Island Loss [43] | FER-2013 | 2018 | 52.52 |
| Covariance Pooling [75] | MS-Celeb-1M | 2018 | 58.14 |
| IPFR [31] | FER-2013 | 2019 | 55.10 |
| RAN (ResNet18) [20] | MS-Celeb-1M | 2020 | 54.19 |
| RAN(ResNet18+VGG16) [20] | MS-Celeb-1M | 2020 | 56.40 |
| LDL-ALSG [70] | AffectNet+RAFDB | 2020 | 56.50 |
| MA-Net (Ours) | FER-2013 | 2020 | **59.40** |

indict that category of contempt exist in many error annotations, which impact model performance severely. And the method of SCN [63] is to address such a problem. Although our MA-Net did not address this problem, primarily, we can obtain the best result from it.

*3) Comparison on RAF-DB:* The comparison with state-of-the-art methods on RAF-DB is presented in Tab. VII. RAF-DB dataset has basic emotion categories and compound categories. Consistent with other methods, we evaluate the effectiveness of MA-Net by recognizing basic emotion categories. The results presented in Tab. VII demonstrate that our MA-Net obtains the highest accuracy of 88.40% on RAF-DB. Even though without pre-training, MA-Net is superior to some state-of-the-art methods with pre-training. It is worth noting that RAN [20] obtain a fantastic performance on RAF-DB by training both on RAF-DB and AffectNet. However, our MA-Net is trained on RAF-DB only, but it is still superior to RAN [20].

*4) Comparison on SFEW:* The recognition accuracy of MA-Net and its comparison with state-of-the-art methods on SFEW are shown in Tab. VIII. In view of the quantity of training set on SFEW is tiny, similar to some state-of-the-art methods, we first pre-train our MA-Net on FER-2013 dataset and then fine-tune our model on SFEW. The results in Tab. VIII demonstrate that MA-Net achieves 59.40% FER accuracy, which significantly outperforms state-of-the-art methods.

### *E. Experiments on Realistic Occlusion and Pose Variation*

To evaluate our method under the real-world scenario, we conduct several experiments on datasets with realistic occlusion and pose variation.

*1) Evaluation of Realistic Occlusion:* To evaluate our method under the occlusion condition, we conduct several experiments on FED-RO, Occlusion-AffectNet, and Occlusion-RAF-DB test set, and the experiment setting is the same as previous work. Fig. 8(a) shows some examples in FED-RO. The comparison with state-of-the-art methods on FED-RO is shown in Tab. IX. Our MA-Net obtains 70.00% accuracy on FED-RO, which is significantly superior to state-of-the-art methods. The comparison with state-of-the-art methods on Occlusion-AffectNet and Occlusion-RAF-DB is shown in Tab. X. We finally achieve the results of 59.59% and 83.65% on two test sets, which is better than 58.50% and 82.72% of RAN [20] respectively. The results on realistic occlusion facial expression datasets demonstrate that our MA-Net has fantastic robustness towards occlusion conditions.

TABLE IX

COMPARISON TO THE STATE-OF-THE-ART RESULTS ON THE FED-RO DATASET

| Methods | Years | Acc.(%) |
|---|---|---|
| VGG-16 [18] | 2015 | 60.15 |
| RseNet-18 [18] | 2016 | 64.25 |
| gACNN [18] | 2019 | 66.50 |
| RAN [20] | 2020 | 67.98 |
| MA-Net (Ours) | 2020 | **70.00** |

TABLE X

COMPARISON TO THE STATE-OF-THE-ART RESULTS ON THE OCCLUSION-AFFECTNET & OCCLUSION-RAF-DB DATASETS

| Datasets | Methods | Acc.(%) |
|---|---|---|
| Occlusion-AffectNet | ResNet-18 [20] | 49.48 |
| | RAN [20] | 58.50 |
| | MA-Net (Ours) | **59.59** |
| Occlusion-RAF-DB | ResNet-18 [20] | 80.19 |
| | RAN [20] | 82.72 |
| | MA-Net (Ours) | **83.65** |

TABLE XI

COMPARISON TO THE STATE-OF-THE-ART RESULTS ON THE POSE-AFFECTNET & POSE-RAF-DB DATASETS

| Datasets | Methods | Pose ($\geqslant 30°$) | Pose ($\geqslant 45°$) |
|---|---|---|---|
| Pose-AffectNet | ResNet-18 [20] | 50.10 | 48.50 |
| | RAN [20] | 53.90 | 53.19 |
| | MA-Net (Ours) | **57.51** | **57.78** |
| Pose-RAF-DB | ResNet-18 [20] | 84.04 | 83.15 |
| | RAN [20] | 86.74 | 85.20 |
| | MA-Net (Ours) | **87.89** | **87.99** |

*2) Evaluation of Realistic Pose Variation:* To evaluate our MA-Net under realistic pose variation condition, we conduct experiments on Pose-AffectNet and Pose-RAF-DB. Fig. 8(b) shows some examples in Pose-AffectNet. The comparison with the state-of-the-art method is shown in Tab. XI, which demonstrates that our MA-Net is significantly superior to RAN [20] on both two test sets. Particularly, comparing the performance between angle larger than 30° and 45°, our MA-Net achieves a tiny reduction of accuracy, indicating that our MA-Net has fantastic robustness to pose variation.

## V. CONCLUSION AND DISCUSSION

In this paper, we propose a global multi-scale and local attention network (MA-Net) to address occlusion and non-frontal pose problems for FER in the wild. The proposed MA-Net is capable of acquiring robust both global and local features, which can address the issues both of occlusion and pose variation well. Specifically, the multi-scale module is employed to fuse features with different receptive fields, which reduces the susceptibility of deeper convolution towards occlusion and variant pose. The local attention module can guide the network to focus on local salient features, which relieves the interference of occlusion and non-frontal pose situations. To verify the effectiveness of the proposed MA-Net under occlusion and pose variation conditions, we carry out experiments on the realistic occlusion and pose variation datasets. The results demonstrate that the proposed MA-Net has strong robustness and outperforms the existing state-of-the-art methods. Moreover, we compare MA-Net with other state-of-the-art methods on some popular datasets, including CAER-S, AffectNet, RAFDB, and SFEW, and MA-Net achieves the best performance on each mentioned datasets.

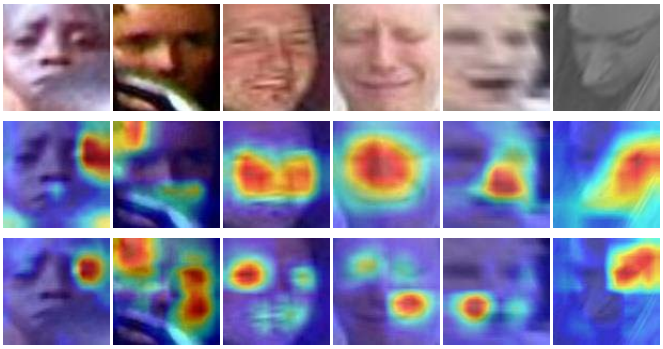


Fig. 9. The failure cases and its class activation mapping (CAM). The images on top are raw data, the CAMs on the middle and the bottom are generated by multi-scale module and local attention module, respectively.

However, the proposed MA-Net will fail in some specific cases. Examples shown in Fig. 9 indicate that the blur is the main problem. For facial emotional images, the blurred images will entail ambiguity of the expression, which usually leads to inconsistent and incorrect labels called noise problems. In our future work, we will attempt to handle the noise problems that exist in facial expression recognition.

## ACKNOWLEDGMENT

The authors would like to thank Prof. Xianwen Xu for English proofreading, and they would also like to thank Kai Wang and Dr. Yong Li for sharing occlusion and pose variation datasets.

Generated by IEEEtran.bst, version: 1.13 (2008/09/30)

## REFERENCES

[1] Z. Duric *et al.*, "Integrating perceptual and cognitive modeling for adaptive and intelligent human-computer interaction," *Proc. IEEE*, vol. 90, no. 7, pp. 1272–1289, Jul. 2002.

[2] M. Jeong and B. C. Ko, "Driver's facial expression recognition in real-time for safe driving," *Sensors*, vol. 18, no. 12, pp. 4270–4288, 2018.

[3] B. Li *et al.*, "A facial affect analysis system for autism spectrum disorder," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 4549–4553.

[4] R. Irani *et al.*, "Spatiotemporal analysis of RGB-D-T facial images for multimodal pain level recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 88–95.

[5] H. Yang, U. Ciftci, and L. Yin, "Facial expression recognition by de-expression residue learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2168–2177.

[6] S. Xie, H. Hu, and Y. Wu, "Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition," *Pattern Recognit.*, vol. 92, pp. 177–191, Aug. 2019.

[7] A. R. Hazourli, A. Djeghri, H. Salam, and A. Othmani, "Deep multi-facial patches aggregation network for facial expression recognition," 2020, *arXiv:2002.09298*. [Online]. Available: http://arxiv.org/abs/2002.09298

[8] Y. Chen, J. Wang, S. Chen, Z. Shi, and J. Cai, "Facial motion prior networks for facial expression recognition," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2019, pp. 1–4.

[9] S. Xie and H. Hu, "Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks," *IEEE Trans. Multimedia*, vol. 21, no. 1, pp. 211–220, Jan. 2019.

[10] Z. Yu, G. Liu, Q. Liu, and J. Deng, "Spatio-temporal convolutional features with nested LSTM for facial expression recognition," *Neurocomputing*, vol. 317, pp. 50–57, Nov. 2018.

[11] S. Wang, H. Shuai, and Q. Liu, "Phase space reconstruction driven spatio-temporal feature learning for dynamic facial expression recognition," *IEEE Trans. Affect. Comput.*, early access, Jul. 7, 2020, doi: 10.1109/TAFFC.2020.3007531.

[12] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. Comput. Soc. Conf. Comput. Vis. Pattern Recognit.-Workshops (CVPRW)*, Jun. 2010, pp. 94–101.

[13] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Apr. 1998, pp. 200–205.

[14] J. Lee, S. Kim, S. Kim, J. Park, and K. Sohn, "Context-aware emotion recognition networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10143–10152.

[15] 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.

[16] 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. 2019.

[17] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Collecting large, richly annotated facial-expression databases from movies," *IEEE MultimediaMag.*, vol. 19, no. 3, pp. 34–41, Jul. 2012.

[18] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using CNN with attention mechanism," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2439–2450, May 2019.

[19] Y. Li, Y. Lu, J. Li, and G. Lu, "Separate loss for basic and compound facial expression recognition in the wild," in *Proc. Asian Conf. Mach. Learn. (ACML)*, 2019, pp. 897–911.

[20] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 4057–4069, 2020.

[21] F. Zhang, T. Zhang, Q. Mao, and C. Xu, "Geometry guided pose-invariant facial expression recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 4445–4460, 2020.

[22] F. Bourel, C. C. Chibelushi, and A. A. Low, "Recognition of facial expressions in the presence of occlusion," in *Proc. Brit. Mach. Vis. Conf.*, 2001, pp. 1–10.

[23] X. Mao, Y. Xue, Z. Li, K. Huang, and S. Lv, "Robust facial expression recognition based on rpca and adaboost," in *Proc. Workshop Image Anal. Multimedia Interact. Services (WIAMIS)*, 2009, pp. 113–116.

[24] B. Jiang and K.-B. Jia, "Research of robust facial expression recognition under facial occlusion condition," in *Proc. Int. Conf. Act. Media Technol. (AMT)*, 2011, pp. 92–100.

[25] Z. Hammal, M. Arguin, and F. Gosselin, "Comparing a novel model based on the transferable belief model with humans during the recognition of partially occluded facial expressions," *J. Vis.*, vol. 9, no. 2, p. 22, 2009.

[26] S. L. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," *IEEE Trans. Affect. Comput.*, vol. 6, no. 1, pp. 1–12, Jan. 2015.

[27] Y. Li, J. Zeng, S. Shan, and X. Chen, "Patch-gated CNN for occlusion-aware facial expression recognition," in *Proc. Int. Conf. Pattern Recognit. (ICPR)*, 2018, pp. 2209–2214.

[28] M. Jampour, V. Lepetit, T. Mauthner, and H. Bischof, "Pose-specific non-linear mappings in feature space towards multiview facial expression recognition," *Image Vis. Comput.*, vol. 58, pp. 38–46, Feb. 2017.

[29] Y.-H. Lai and S.-H. Lai, "Emotion-preserving representation learning via generative adversarial network for multi-view facial expression recognition," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 263–270.

[30] F. Zhang, T. Zhang, Q. Mao, and C. Xu, "Joint pose and expression modeling for facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3359–3368.

[31] C. Wang, S. Wang, and G. Liang, "Identity- and pose-robust facial expression recognition through adversarial feature learning," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 238–246.

[32] G. Yovel and B. Duchaine, "Specialized face perception mechanisms extract both part and spacing information: Evidence from developmental prosopagnosia," *J. Cognit. Neurosci.*, vol. 18, no. 4, pp. 580–593, Apr. 2006.

[33] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 818–833.

[34] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021.

[35] P. C. Ng, "SIFT: Predicting amino acid changes that affect protein function," *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3812–3814, Jul. 2003.

[36] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 886–893.

[37] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced Fisher linear discriminant model for face recognition," *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 467–476, Apr. 2002.

[38] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas, "Learning active facial patches for expression analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 2562–2569.

[39] P. Yang, Q. Liu, and D. N. Metaxas, "RankBoost with L1 regularization for facial expression recognition and intensity estimation," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 1018–1025.

[40] H. Ding, S. K. Zhou, and R. Chellappa, "FaceNet2ExpNet: Regularizing a deep face recognition net for expression recognition," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2017, pp. 118–126.

[41] W. Hayale, P. Negi, and M. Mahoor, "Facial expression recognition using deep Siamese neural networks with a supervised loss function," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–7.

[42] H. Siqueira, S. Magg, and S. Wermter, "Efficient facial feature learning with wide ensemble-based convolutional neural networks," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2020, pp. 5800–5809.

[43] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly, and Y. Tong, "Island loss for learning discriminative features in facial expression recognition," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 302–309.

[44] G. Zeng, J. Zhou, X. Jia, W. Xie, and L. Shen, "Hand-crafted feature guided deep learning for facial expression recognition," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 423–430.

[45] B. Pan, S. Wang, and B. Xia, "Occluded facial expression recognition enhanced through privileged information," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 566–573.

[46] L. Zhong, Q. Liu, P. Yang, J. Huang, and D. N. Metaxas, "Learning multiscale active facial patches for expression analysis," *IEEE Trans. Cybern.*, vol. 45, no. 8, pp. 1499–1510, Aug. 2015.

[47] Y. Liu, X. Yuan, X. Gong, Z. Xie, F. Fang, and Z. Luo, "Conditional convolution neural network enhanced random forest for facial expression recognition," *Pattern Recognit.*, vol. 84, pp. 251–261, Dec. 2018.

[48] L. Zhong, C. Bai, J. Li, T. Chen, S. Li, and Y. Liu, "A graph-structured representation with brnn for static-based facial expression recognition," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–5.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[50] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2015, pp. 91–99.

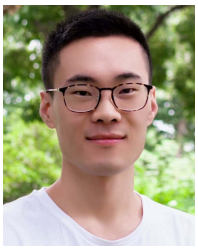
[51] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 21–37.

[52] M. Najibi, P. Samangouei, R. Chellappa, and L. S. Davis, "SSH: Single stage headless face detector," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4875–4884.

[53] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (And a dataset of 230,000 3d facial landmarks)," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1021–1030.

[54] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[55] Y. Fan, J. C. Lam, and V. O. Li, "Video-based emotion recognition using deeply-supervised neural networks," in *Proc. ACM Int. Conf. Multimodal Interact. (ICMI)*, 2018, pp. 584–588.
[56] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
[57] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.
[58] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.
[59] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.
[60] J. Cai *et al.*, "Feature-level and model-level audiovisual fusion for emotion recognition in the wild," in *Proc. IEEE Conf. Multimedia Inf. Process. Retr. (MIPR)*, Mar. 2019, pp. 443–448.
[61] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.
[62] J. Zeng, S. Shan, and X. Chen, "Facial expression recognition with inconsistently annotated datasets," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 222–237.
[63] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6897–6906.
[64] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-shot multi-level face localisation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5203–5212.
[65] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2019, pp. 8026–8037.
[66] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 87–102.
[67] I. J. Goodfellow *et al.*, "Challenges in representation learning: A report on three machine learning contests," *Neural Netw.*, vol. 64, pp. 59–63, Apr. 2015.
[68] D. Kollias, S. Cheng, E. Ververas, I. Kotsia, and S. Zafeiriou, "Deep neural network augmentation: Generating faces for affect analysis," *Int. J. Comput. Vis.*, vol. 128, no. 5, pp. 1455–1484, May 2020.
[69] Y. Fu, X. Wu, X. Li, Z. Pan, and D. Luo, "Semantic neighborhood-aware deep facial expression recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 6535–6548, 2020.
[70] S. Chen, J. Wang, Y. Chen, Z. Shi, X. Geng, and Y. Rui, "Label distribution learning on auxiliary label space graphs for facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13984–13993.
[71] C. Hewitt and H. Gunes, "CNN-based facial affect analysis on mobile devices," 2018, *arXiv:1807.08775*. [Online]. Available: http://arxiv.org/abs/1807.08775
[72] X. Liu, B. V. Kumar, J. You, and P. Jia, "Adaptive deep metric learning for identity-aware facial expression recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 20–29.
[73] Z. Meng, P. Liu, J. Cai, S. Han, and Y. Tong, "Identity-aware convolutional neural network for facial expression recognition," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2017, pp. 558–565.
[74] N. Otberdout, A. Kacem, M. Daoudi, L. Ballihi, and S. Berretti, "Deep covariance descriptors for facial expression recognition," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2018, pp. 1–13.
[75] D. Acharya, Z. Huang, D. P. Paudel, and L. Van Gool, "Covariance pooling for facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 367–374.



**Zengqun Zhao** (Graduate Student Member, IEEE) received the B.S. degree in electronic information engineering from the Shaanxi University of Technology, Hanzhong, China, in 2018. He is currently pursuing the M.S. degree in control science and engineering with the School of Automation, Nanjing University of Information Science and Technology, Nanjing, China. His research interests include computer vision and deep learning.

**Qingshan Liu** (Senior Member, IEEE) received the M.S. degree from Southeast University, Nanjing, China, in 2000, and the Ph.D. degree from the National Laboratory of Pattern Recognition, Chinese Academy of Sciences, Beijing, China, in 2003. He is currently a Professor with the School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing. His research interests include pattern recognition, and image and video analysis.

**Shanmin Wang** received the B.S. degree in automation and the M.S. degree in control science and engineering from the Nanjing University of Information Science and Technology, Nanjing, China, in 2017 and 2020, respectively. She is currently pursuing the Ph.D. degree in computer science and engineering with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing. Her research interests include computer vision and deep learning.