IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 1, 2020, accepted February 17, 2020, date of publication February 24, 2020, date of current version March 3, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2975913

Weakly Supervised Local-Global Attention Network for Facial Expression Recognition

HAIFENG ZHANG[1], WEN SU[3], AND ZENGFU WANG[1,2], (Member, IEEE)
[1]Department of Automation, University of Science and Technology of China, Hefei 230022, China
[2]Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei 230031, China
[3]Faculty of Mechanical Engineering and Automation, Zhejiang Sci-Tech University, Hangzhou 310018, China

Corresponding author: Zengfu Wang (zfwang@ustc.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61472393.

ABSTRACT Combining global and local features is an essential solution to improve discriminative performances in facial expression recognition tasks. The limitations of existing methods are that they cannot extract crucial local features and ignore the complementary effects of local and global features. To address these problems, this paper proposes a Weakly Supervised Local-Global Attention Network (WS-LGAN), which uses the attention mechanism to deal with part location and feature fusion problems. Firstly, an Attention Map Generator is designed to get a set of attention maps under weak supervision. It mimics the attention mechanism of human brain and quickly finds the local regions-of-interest. Secondly, bilinear attention pooling is employed to generate and refine local features based on attention maps. Thirdly, a building block called Selective Feature Unit is designed. It allows adaptive weighted fusion of global and local features before making classification. In WS-LGAN, global and local features represent expressions from different aspects. Compared with methods relying on single type of feature, it benefits from local-global complementary advantages. Additionally, contrastive loss is introduced for both local and global features to increase inter-class dispersion and intra-class compactness under different granularities. Experiments on three popular facial expression datasets, including two lab-controlled facial expression datasets and one real-world facial expression dataset show that WS-LGAN achieves state-of-the-art performance, which demonstrates our superiority in facial expression recognition.

INDEX TERMS Facial expression recognition, weak supervision, attention mechanism, local features, global features.

I. INTRODUCTION

Facial expression is a fundamental manner of transporting human emotions and takes on a significant part in our daily communication. Facial expression recognition is a complex but interesting problem, and finds its extensive applications in fatigue surveillance [1], human-machine interaction [2], patient care [3], neuromarketing [4] and interactive games [5] etc. Thus, facial expression recognition has received substantial attention among the researchers in computer vision, affective computing and human computer interaction fields.

Despite great success has been achieved in recent years [6]–[8], accurate facial expression recognition is still challenging. It is mainly due to the complexity and variability of facial expressions. We summarize the obstacles as follows: (a) High intra-class variances. Expressions of the same class may vary from one person to another. It is influenced by factors such as age, gender, race, cultural background and other person-specific characteristics. (b) Low inter-class variances. Expressions belonging to different categories may be similar except for some minor differences. For example, sadness and anger sharing traits of uniformity across some facial regions. The main difference of sadness and anger only lies in the corners of the mouth. The existing methods for overcoming the above two obstacles can be divided into two categories. One category is non-part based method which focuses on learning global representation, while the other is part-based based method which pays more attention to extract partial discriminative features.

For the first category, several works propose novel loss layers to replace or assist the supervision of the softmax loss for more discriminative features. Inspired by the center

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad Ayoub Khan.

loss [9] and the triplet loss [10], some variations such as, island loss [11], locality-preserving loss [12] and (N+M)-tuples cluster loss [13] are designed for facial expression recognition. They require projecting the features to another space in which inter-class discrimination and intra-class similarity are enhanced. However, these methods usually extract features from the holistic facial image. It ignores fine-grained information in local facial regions.

For the second category, an essential prerequisite of learning discriminative part features is that parts should be accurately located. Since facial expression datasets usually have only image-level expression labels, we cannot have an extra part annotation, such as pixel-level segmentation labels or bounding boxes. Some part-based methods crop facial expression image into patches and try to learn local representation of them. For instance, [14]–[16] divide facial image into non-overlapping or overlapping patches. Features extracted from selected patches or all patches can highlight some details about local facial regions. Although their results are encouraging, they still have shortcomings. Firstly, selected facial patches may vary with the training data, it is difficult to conceive a generic system. Secondly, a large number of candidate patches make the training of the model time-consuming and computationally intensive. Thirdly, manually defined patches may not be optimal for the final classification tasks. Some patches have no influence on expression or even have a negative impact on facial expression recognition.

Moreover, we may lose some complementary information if we only concentrate on local features. Attributes such as age, gender, race and other person-specific characteristics, provided by the holistic facial image, can also affect expressions significantly. Thus, methods based on either localized or global features alone will ignore their joint benefit and mutual complementary effects.

In fact, when humans try to conduct an object recognition process, we firstly obtain the global description. Then our attention orientates rapidly toward salient regions where facial details will be filled out [17]. Besides, studies in [18]–[20] have provided a prior knowledge that much of expressional clues come from salient facial regions, such as the regions around mouth and eyes. Conversely, other parts have little impact on facial expression recognition, such as ears and hair. Motivated by the process and the prior knowledge, we propose a Weakly Supervised Local-Global Attention Network (WS-LGAN) to learn global representations and, at the same time, learn local features around eyes and mouth to facilitate local-enhanced facial expression recognition. The pipeline of our proposed method is illustrated in Figure 2. Different from previous part-based methods, we mimic the way humans recognize facial expressions. Attention mechanism is introduced to guide our network to perceive crucial local regions autonomously. The location of crucial facial region is designated by attention maps which are generated by Attention Map Generator (AMG). Attention maps learning is only weakly supervised by image-level label. Therefore, the lack of part annotations in facial expression dataset is well solved. We show some samples generated by AMG in Figure 1. Inspired by [28], we propose a bilinear attention pooling. Based on the attention maps, the local features are extracted and refined by combining the attention maps and the global feature through bilinear attention pooling. In the same time, we integrate global features to introduce complementary information displayed in holistic facial images. Local features and global features are fused with adaptive weights through a Selective Feature Unit (SFU). Furthermore, we develop similarity metric for local features and combine it with classification errors. They assimilate local features extracted from same category and discriminate local features extracted from different categories. As a result, the intra-class variations of local features are reduced, while the inter-class differences of local features are increased. Similarly, the similarity metric is also utilized in global features.

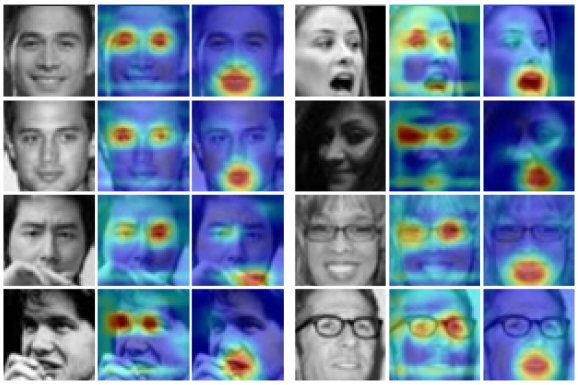


FIGURE 1. Attention maps that indicate crucial facial regions. For a given input image (left), eye-related attention map (center) shows the location of eyes and mouth-related attention map (right) shows the location of mouth. Best viewed in color.

In summary, the main contributions of this work are as follows:

(1) *Weakly supervised model for local feature extraction:* We propose a local feature extraction method that explicitly considers and extracts region-specific local features, including features around eyes and mouth. Most previous models obtain local representation relies on segmented expressional image patches [14]–[16]. A large number of candidate facial image patches makes the model inefficient, while expression-independent patches hinder the training of the model. In addition, the position and scale of crucial local regions also change with the input image, it is difficult to determine the size of the patch. Different from these methods, we propose handling local feature extraction by directly locating the crucial regions and extracting the corresponding features. Specifically, our method trains the AMG under weak supervision to generate attention maps that strongly indicate the locations of the eyes and mouth. Based on these attention maps, bilinear attention pooling is proposed to generate and refine local features. Such local feature extraction method has two advantages comparing with previous methods. Firstly, we

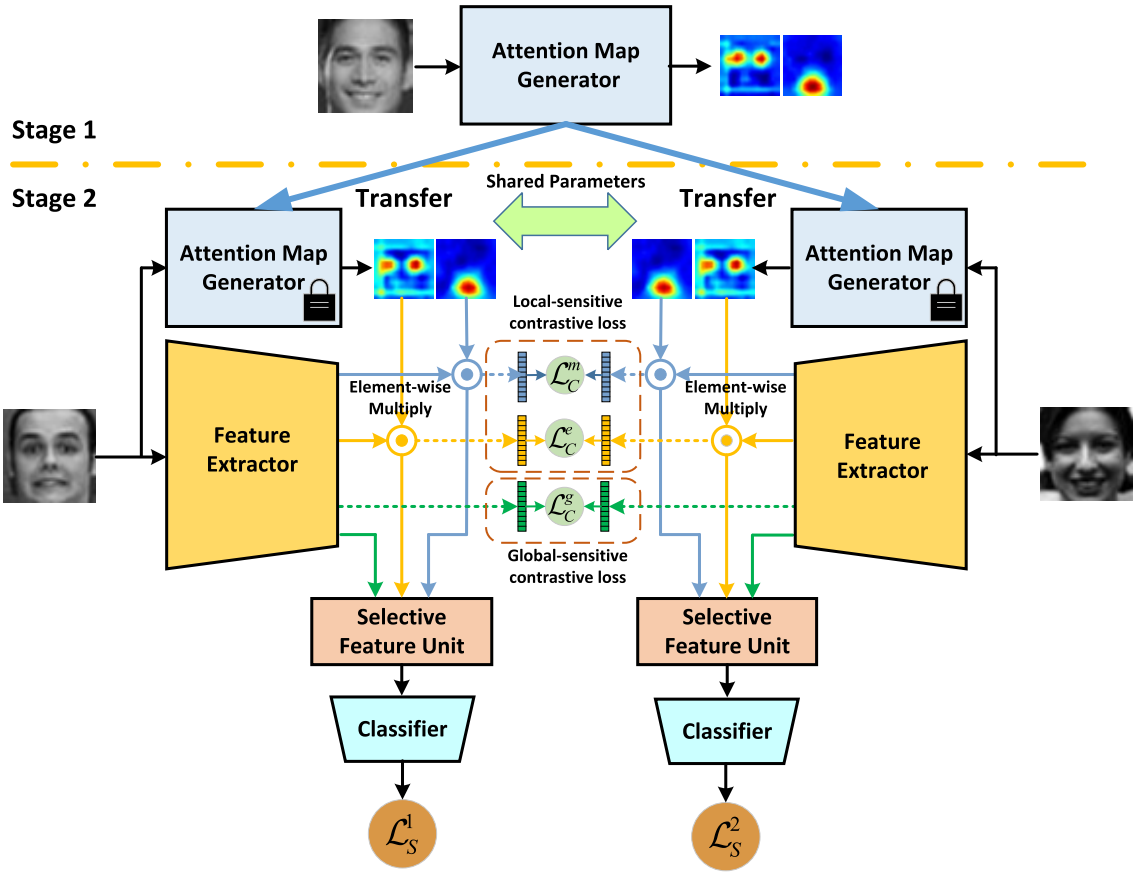


FIGURE 2. Overview of the proposed model. We train Attention Map Generator in the first stage and transfer it to the second stage with weight fixed. In the second stage, a model consists of two identical CNN streams, whose weights are shared, is proposed to extract and fuse both local features and global features. Feature Extractor extracts features directly from the holistic facial image. Without any additional processing, the feature extracted by Feature Extractor is the global feature. Local features are extracted and refined by combining the attention maps and the global feature. Selective Feature Unit fuses all features with adaptive weights. During testing, only one input image is fed into one CNN stream. Best viewed in color.

markedly reduce the number of candidate local regions, expression-irrelevant local regions are discarded. Secondly, changes in the position and size of the mouth and eyes regions will not affect the settings of the model. There is no need to consider the size of the patches. Besides, weak supervision allows us to overcome the limitation of no part annotations in facial expression dataset, resulting in a concise feature extraction.

(2) *Adaptive weighted local-global fusion:* We formulate a SFU to fuse local and global features. The local features focus on extracting region-specific fine-grained information, while the global features concentrate on representing the integrity of the expression. Compared with methods relying on either local or global representation alone, the joint use of them makes the model can benefit from their complementary advantages. Previous method [14] aggregates global feature and all locals by concatenate fusion. However, such a fusion method treats all features uniformly, while our proposed SFU enables an adaptive weighted fusion. Specifically, the SFU has two effects on the features. On one hand, it weights the specific gravity of features extracted from region around eyes, region around mouth, and the holistic face region, respectively. On the other hand, it weights different semantic information and finds the most meaningful feature within each region.

(3) *Multi-granular similarity metrics:* We extend metric learning to both local features and global features to increase inter-class differences as well as decrease intra-class variations at different granularities. Previous methods [11]–[13], [22] employ similarity metrics only on global representation, and thus fine-grained features are not well learned. Part-based methods [14]–[16] which extract local features through patches are difficult to metric learning on local features because the facial pose changes in each image. Among multiple patches, patches at the same position may correspond to different facial part. In our method, we have located the eyes and mouth regions on each image. This is equivalent to ''aligning'' the mouth and eyes on each image. Under the ''alignment'', we can perform metric learning on mouth-related features and eyes-related features, respectively. Therefore, we propose local-sensitive contrastive loss for eyes-related and mouth-related features. We are able to

make full use of local-sensitive contrastive loss and global-sensitive contrastive loss simultaneously.

(4) *Competitive experimental results:* To demonstrate the superiority of our proposed method, we employ experiments on lab-controlled facial expression datasets (e.g., CK+, Oulu-CASIA) and real-world facial expression dataset (namely, RAF-DB). Our facial expression recognition solution achieves state-of-the-art results on CK+, Oulu-CASIA and RAF-DB with accuracies of 98.06%, 88.26% and 85.07%, respectively.

The remainder of this paper is organized as following. Section 2 reviews related work on facial expression recognition. Section 3 details the proposed Weakly Supervised Local-Global Attention Network. In section 4, we show the experimental results and evaluate the performance of the WS-LGAN. In Section 5, we conclude the paper and give some remarks.

II. RELATED WORK

Researchers have long acknowledged that facial expression recognition struggles when coupled with inter-subject variations. The visual differences among categories or instances are subtle. They are easily overwhelmed by other factors. Existing methods that concentrate on these problems can be classified into two categories: non part-based methods and part-based methods.

A. NON PART-BASED METHODS

An immensely popular recent approach is to enhance the discriminative power of the deeply learned features by proposing new loss functions. These methods aim to obtain representation with compact intra-class variations and separable inter-class differences. Cai *et al.* [11] propose an island loss, which can penalise the distance between deep features and their corresponding class centers as well as increase the pairwise distances between different class centers simultaneously. Li *et al.* [12] propose a deep locality-preserving CNN, which preserves the locality proximity by minimizing the distance to the K-nearest neighbors within the same class.

Besides, some works attempt to make the network disentangle the identity and the expression by either performing multi-signal supervision or using Generative Adversarial Network. Meng *et al.* [22] propose a model that contains two identical sub-CNNs. One stream learns expression-discriminative features, and the other stream learns identity-related features for identity-invariant expression recognition. Liu *et al.* [13] propose (N+M)-tuples cluster loss with the supervision of identity and expression labels to alleviate the difficulty of anchor selection and threshold validation in the triplet loss for identity-invariant facial expression recognition. Hui *et al.* [23] learn facial expressions by extracting the expressive component through a de-expression procedure. Given a facial image with arbitrary expressions, its corresponding neutral expression is generated by the trained generative model. Through the procedure, the identity information of a subject remains unchanged while the expressive component is removed. The expressive component is used to make facial expression recognition.

B. PART-BASED METHODS

Studies in psychology show that most of the descriptive facial features of expressions are located in several crucial regions. Therefore, extract local features from facial image has attracted the attention of some researchers. Some part-based methods have been proposed. Happy and Routray [15] propose a framework by using appearance features of selected facial patches. They select different facial patches as salient for different expressions. Liu *et al.* [16] model a system named boosted deep belief network to classify different expressions. They divide expressional images into patches. Some patches with high discriminative power are selected and combined to train a strong classifier. 3DCNN-DAP [24] incorporate a deformable parts learning component into the 3D CNN framework, which can detect specific facial action parts under the structured spatial constraints, and obtain the discriminative part-based representation simultaneously.

Recently, some methods have demonstrated that integrating local and global features can improve the performance of facial expression recognition. For instance, Xie and Hu [14] propose a convolutional neural network with two branches. One branch extracts local features from uniform image patches while the other extracts global features from the holistic expressional image. Global features and all local features are concatenated for final expression classification.

III. PROPOSED METHOD

A. OVERVIEW

Our method aims to mine discriminative parts of the face via object localization and extract local features through the mined parts. Besides, we fuse local and global features to utilize their complementary advantages jointly in coping with local detail loss and emphasizing global integrity. Since facial expression datasets do not have labeled part locations, we formulate part localization problem in a weakly supervised manner by introducing a facial attributes dataset. We decompose our pipeline into two stages, as shown in Figure 2. We train the AMG on facial attributes dataset in the first stage. Then the well-trained AMG is transferred to facial expression datasets with weight fixed. Local and global features are jointly learned through a deep CNN framework in the second stage. The second stage consists of two identical CNN streams whose weights are shared. Each CNN stream contains four sub-parts: AMG, Feature Extractor (FE), SFU and Classifier. The AMG generates attention maps by calculating weighted feature map in the binary classification network. The attention maps can highlight the regions around eyes and mouth. FE extracts features directly from the holistic facial image. Without any additional processing, the feature extracted by FE is the global feature. Local features are extracted and refined by combining the attention maps and

the global feature. The backbone of each branch in the AMG is a variant of Densenet [25]. It consists of 3 dense blocks and 2 transition layers. The dense block contains 6, 12 and 24 dense layers, respectively. Due to the limited images in facial expression datasets, we use the backbone as the FE after reducing the number of dense layers to 6 for each dense block. The SFU aims to fuse the global features with the local features. Classifier is a softmax classifier for the final expression classification. During training, the model takes a pair of facial expression images as input. We optimize the parameters by simultaneously minimizing the classification errors, local-sensitive contrastive loss and global-sensitive contrastive loss. During testing, an image is fed into one CNN stream, and predictions are generated based on both the local and the global features.

B. ATTENTION MAP GENERATOR

Attention map is a weight map where crucial regions have higher values. We use it to locate crucial facial regions. In general, the direct way of locating a region is to use an image and its pixel-wise segmentation as input and target respectively, such as facial parsing. However, it requires label maps with pixel-wise annotations, which are expensive to collect. Pixel-level image processing can be time-consuming and computationally expensive. More importantly, facial expressions are generated by contracting facial muscles around facial organs. The result of pixel segmentation is too fine to focus on the areas around these organs that contain abundant apparent features. An alternative approach is weakly supervised object localization. Zhou *et al.* [26] enable the classification network to have remarkable localization ability despite being trained on image-level labels. It can be applied to a variety of computer vision tasks for fast and accurate localization. Inspired by them, we designed our AGM.

Facial expression datasets usually have only expression labels, while the image in the CelebA dataset [27] is labeled with 40 facial attributes. Some attributes can guide the AMG training to locate crucial regions. Since we only focus on the regions that related to facial expressions, eyes and mouth related facial attributes for each image are selected and divided into two groups based on their respective facial parts. For instance, bushy eyebrows, arched eyebrows, narrow eyes, eyeglasses are grouped together, as all of them are related to eyes. The grouped attributes are summarized in Table 1.

TABLE 1. Facial attributes grouping.

Part	Attributes
Eyes	Bushy eyebrows, Arched eyebrows, Narrow eyes, Eyeglasses
Mouth	Big lips, Mouth slightly open, Smiling

The AMG consists of two branches that locate the regions around the eyes and mouth respectively. Figure 3 takes the eyes-related branch as an example. If the image does not contain any eyes-related attributes listed in Table 1, let's take it as a negative example, otherwise, as a positive example of the eyes. We use them to train eyes-related branch. Fully convolutional layers and global average pooling (GAP) are used to generate features for classification. The predicted class score is mapped back to the previous convolutional layer to generate the attention maps. As illustrated in Figure 3, GAP outputs the spatial average of the feature map of each unit at the last convolutional layer. A weighted sum of these values is used to generate the final output. We back the weights of the output layer to the convolutional features and compute a weighted sum of the feature maps to obtain our attention maps. We normalize the attention map so that all values fall in the range [0, 1].

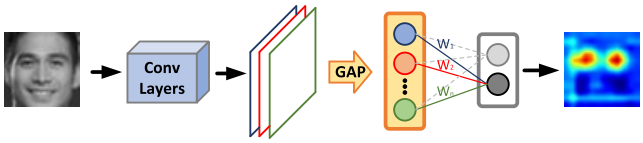


FIGURE 3. One branch in the Attention Map Generator. Conv Layers denotes the backbone. GAP denotes global average pooling.

Weakly supervised part location allows us to overcome the limitation of no part annotations in facial expression dataset. Figure 1 illustrates the effect of attention maps outputted using the AMG. The regions around eyes and mouth are highlighted. After we trained the AMG module on CelebA dataset, we transfer it to the facial expression datasets. In the second stage, the AMG is frozen.

C. LOCAL FEATURE REFINEMENT

Bilinear pooling has been proved to be effective in extracting fine-grained features [28]. It is developed to localize distinct object parts and model the appearance conditioned on their detected locations, but it cannot obtain the details of specific regions. We propose bilinear attention pooling which naturally combines the attention maps to solve this problem. Local feature refinement for two crucial local regions is implemented through bilinear attention pooling. Besides, contrastive loss is utilized for each local feature to learn a similarity metric for image pairs. It makes sure the local features extracted from samples of the same expression have similar representations. On the contrary, those of different expressions are faraway in feature space. As illustrated in Figure 4, we take the refinement of eyes-related features as an example. The process of mouth-related features refinement is the same as that of eye-related features.

1) BILINEAR ATTENTION POOLING

Firstly, well-trained AMG with fixed weights is used to generate attention maps $A_e \in R^{1\times H\times W}$ (eyes-related attention maps) and $A_m \in R^{1\times H\times W}$ (mouth-related attention maps) respectively. Note that, A_e and A_m have the same map size $H \times W$. 1 is the number of channel, R is the set of real number. Then, we element-wise multiplies each channel in feature maps $F_g \in R^{C\times H\times W}$ (C is the number of channels) by attention maps A_e and A_m, as shown in Eq.1:

$$F_e^i = A_e \odot F_g^i \tag{1}$$

$$F_m^i = A_m \odot F_g^i \tag{2}$$

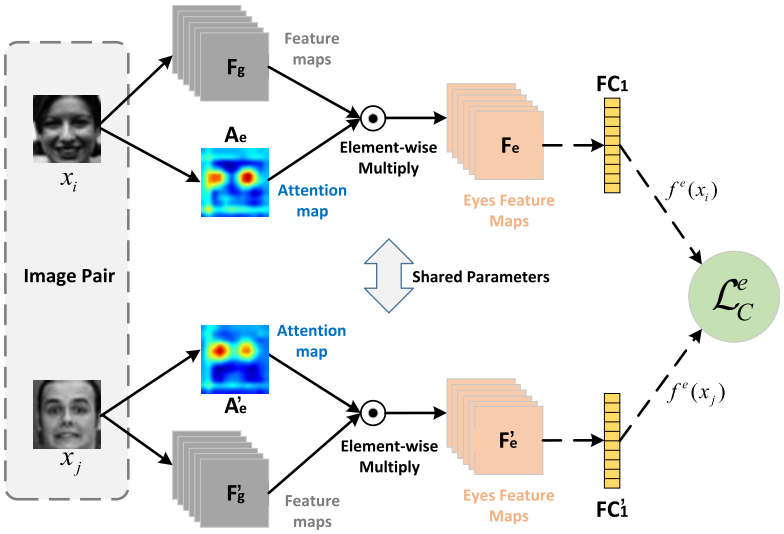


FIGURE 4. The process of refining eyes-related features. Eyes feature maps are generated by element-wise multiplying attention maps with feature maps. Local-sensitive contrastive loss is employed on the representation generated by two component networks with parameters shared.

where $\odot$ indicates element-wise multiplication for two tensors. Feature maps F_g are extracted by the FE from the holistic facial image. $F_g^i \in R^{1\times H\times W}$ is the ith channel in F_g. After the element-wise multiplication, F_e^i and F_m^i represent the ith feature map of eyes and mouth, respectively. We concatenate the F_e^i across the channel to obtain F_e, and concatenate the F_m^i across the channel to obtain F_m. $F_e \in R^{C\times H\times W}$ and $F_m \in R^{C\times H\times W}$ reflect the feature maps of eyes and mouth, respectively.

Bilinear attention pooling explicitly define two streams to locate and extract features respectively. We treat the AMG branch as the dorsal stream in the human visual cortex, which processes the object's spatial location and the FE branch as the ventral stream in the human visual cortex, which performs object recognition. The bilinear attention pooling bridges the relationship between appearance models and part locating models. It provides a solution for local feature extraction.

2) LOCAL-SENSITIVE CONTRASTIVE LOSS

To reduce the intra-class variations and enlarge the inter-class differences at finer granularity. Local-sensitive contrastive loss L_C^e and L_C^m are designed for the eyes-related features and mouth-related features respectively. As illustrated in Figure 4, we introduce an auxiliary fully connected (FC) layer to represent the eyes-related features. L_C^e pulls the eyes-related features extracted from samples of the same expression towards each other, while push the eyes-related features extracted from samples of different expressions away from each other. We adopt the loss function based on the squared Euclidean distance, which is denoted as:

$$L_C^e(\theta_{ij}, f^e(x_i), f^e(x_j)) = \begin{cases} \frac{1}{2}(\|f^e(x_i) - f^e(x_j)\|_2^2 & if: \theta_{ij} = 1 \\ \frac{1}{2}\max(0, \delta^e - \|f^e(x_i) - f^e(x_j)\|_2)^2 & if: \theta_{ij} = 0 \end{cases} \quad (3)$$

where x_i and x_j are a pair of training images, and $f^e(x_i)$ and $f^e(x_j)$ are their corresponding eyes-related feature vectors. When x_i and x_j have the same expression label, $\theta_{ij} = 1$. Otherwise, $\theta_{ij} = 0$. δ^e is the size of the margin which determines how much dissimilar pairs contribute to the loss function. In our experiment, δ^e is set to 10 empirically. The contrastive loss L_C^m of mouth-related features is defined similar to L_C^e.

D. LOCAL-GLOBAL FUSION

We define a global-sensitive contrastive loss to extract more discriminative global features. Besides, to enable WS-LGAN to infer image categories from both local details and global context cues concurrently, we propose a SFU to fuse local and global features.

1) GLOBAL-SENSITIVE CONTRASTIVE LOSS

Global-sensitive contrastive loss L_C^g is designed to reduce the intra-class variations and enlarge the inter-class differences globally. Global feature maps $F_g \in R^{C\times H\times W}$ are extracted by the FE from holistic facial image directly without any additional processing. The global feature vector that used to calculate the loss is obtained by inputting F_g into the FC layer. Similar to the local-sensitive contrastive loss, L_C^g is defined as following:

$$L_C^g(\theta_{ij}, f^g(x_i), f^g(x_j)) = \begin{cases} \frac{1}{2}(\|f^g(x_i) - f^g(x_j)\|_2^2 & if: \theta_{ij} = 1 \\ \frac{1}{2}\max(0, \delta^g - \|f^g(x_i) - f^g(x_j)\|_2)^2 & if: \theta_{ij} = 0 \end{cases} \quad (4)$$

where $f^g(x_i)$ and $f^g(x_j)$ are global feature vectors for a pair of training samples. When x_i and x_j have the same expression label, $\theta_{ij} = 1$. Otherwise, $\theta_{ij} = 0$. δ^g is the size of the margin

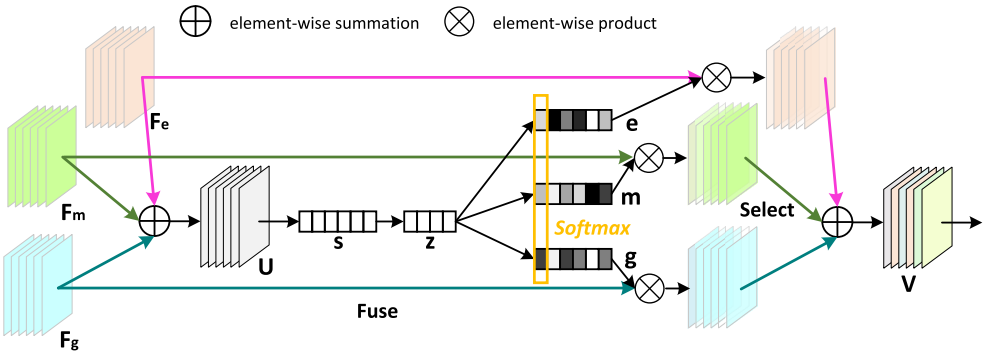


FIGURE 5. Selective Feature Unit.

which determines how much dissimilar pairs contribute to the loss function. It is set to 10 empirically.

2) SELECTIVE FEATURE UNIT

The SFU is inspired by Selective Kernel (SK) convolution (SK convolution) [29], which proposes a dynamic selection mechanism in CNNs that allows each neuron to adaptively adjust its receptive field size based on multiple scales of input information. The main difference between the SFU and the SK convolution is motivation. The SFU is designed to learn an adaptive weighted fusion of features. It models the complementarity of information between local and global features, while the SK convolution aims to address the adaptive changing of receptive fields size. Specifically, the SFU consists of two key schemes: fuse and select as illustrated in Figure 5.

a: FUSE

To compute the adaptive weights for F_e, F_m and F_g, we use gates to control the information flows from multiple branches carrying features extracted from different regions into the next layer. The gates integrate information from all branches. We first obtain the hybrid representation $U = F_e + F_m + F_g$, ($U \in R^{C\times H\times W}$) from three branches via an element-wise summation. Then, we adopt average-pooling to squeeze the spatial dimension of the hybrid representation and generate a channel-wise statistics $s \in R^C$. Further, s is fed to a fully connected layer with ReLU function and Batch Normalization to reduce the dimensionality. It generates a compact feature $z \in R^{d\times 1}$ to guide the precise and adaptive selection. Here, d represents the length of the vector z. The size of d is determined by the number of output neurons of the fully connected layer with s as input. In our experiment, $d = C/16$.

b: SELECT

A soft attention across channels is used to adaptively select three different spatial feature descriptors: F_e, F_m and F_g, which are guided by the compact feature descriptor z. Specifically, a softmax operator is applied on the channel-wise digits:

$$e_c = \frac{e^{E_c z}}{e^{E_c z} + e^{M_c z} + e^{G_c z}} \tag{5}$$

$$m_c = \frac{e^{M_c z}}{e^{E_c z} + e^{M_c z} + e^{G_c z}} \tag{6}$$

$$g_c = \frac{e^{G_c z}}{e^{E_c z} + e^{M_c z} + e^{G_c z}} \tag{7}$$

where the matrix $E, M, G \in R^{C\times d}$. E, M and G are three matrices that can be learned by the model. Each row of E, M and G can be used to calculate each element of e, m and g, respectively. e, m, g denote the soft attention vectors for F_e, F_m and F_g, respectively. $E_c \in R^{1\times d}$ is the cth row of E and e_c is the cth element of e, likewise M_c and m_c, G_c and g_c. The final representation V is obtained through the attention weights on three different spatial feature descriptors:

$$V_c = e_c \cdot F_{e_c} + m_c \cdot F_{m_c} + g_c \cdot F_{g_c} \tag{8}$$

$$e_c + m_c + g_c = 1 \tag{9}$$

where $V = [V_1, V_2, \ldots, V_C]$, $V_c \in R^{H\times W}$.

The SFU plays two different roles in our method. Firstly, features from different spatial descriptors (F_e, F_m and F_g) make different contributions to facial expression recognition task. Between each spatial descriptor, the SFU plays the role of weighting the specific gravity of different spatial descriptors. It estimates the weight adaptively to denote the importance of each spatial descriptor and makes a reasonable trade-off and selection among them. Besides, Besides, the adaptive weight can prevent the model from being sensitive to the performance of AMG. Secondly, as each channel of a feature map is considered as a feature detector, each channel represents features with different semantic information. Within each spatial feature descriptor, the SFU plays the role of weighted different semantic and finding the most meaningful feature from each spatial descriptor.

E. TOTAL LOSS

After the SFU, the final fused representation is fed into the Classifier, which is performed using the softmax classifier. Softmax loss that calculates the classification error is used

on the end of each component network to ensure the learned features are meaningful for expression recognition. The total loss of the proposed WS-LGAN is:

$$L_{total} = \lambda_1 L_C^e + \lambda_2 L_C^m + \lambda_3 L_C^g + \lambda_4 L_S^1 + \lambda_5 L_S^2 \quad (10)$$

where $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6\}$ are the weight of each loss. L_C^e and L_C^m are the local-sensitive contrastive losses that correspond to eyes-related features and mouth-related features, respectively. L_C^g is the global-sensitive contrastive loss, L_S^1 and L_S^2 are the final classification errors.

IV. EXPERIMENTS

In this section, we conduct comprehensive experiments to verify the effectiveness of WS-LGAN. We compare our method with the state-of-the-art methods on three popular facial expression datasets. To demonstrate the effectiveness of our proposed components, we also employ a series of ablation studies. For representing the results of the AMG on facia expression datasets, we visualize attention maps intuitively.

A. DATASETS

1) CELEBA DATASET

The CelebA dataset is an additional dataset we introduced. Note that, we only use CelebA dataset [27] to train the AMG. The dataset contains 202,599 web-based images. Every image are labeled with 40 facial attributes. We select 7 facial attributes for each image and divide the attributes into two groups based on their respective facial parts as shown in Table 1. We randomly select 30,000 (The ratio of positive and negative samples is 1:1) images to train the eyes-related branch and select 3,000 images for validation. For the training of mouth-related branches, we use the same configuration.

2) CK+ DATASET

The CK+ [30] dataset is a lab-controlled dataset which consists of 593 facial expression sequences collected from 123 different subjects. It is an extended version of Cohn-Kanade (CK) dataset [31]. Its subjects range from 18 to 30 years old, most of whom are female. Each sequence starts with a neutral facial expression and ends with a peak facial expression. Among these sequences, only 327 sequences from 118 subjects are annotated with seven expressions, i.e. Anger (An), Disgust (Di), Fear (Fe), Happiness (Ha), Sadness (Sa), Surprise (Su) and Contempt (Co). As a general procedure [11], [13], [14], [16], [21]–[23], [46], the last three frames of each sequence are used for training and testing. Thus, CK+ contains 981 images for our experiments.,

3) OULU-CASIA DATASET

The Oulu-CASIA [33] dataset is a lab-controlled dataset which contains 480 facial expression sequences collected from 80 different subjects aged between 23 and 58 years old. Similar to the CK+, each sequence begins with a neutral expression and ends with a peak expression. All sequences have been labelled with six expressions: anger, disgust, fear, happiness, sadness or surprise. The last three frames of each sequence are selected for training and testing [11], [21], [23], [32]. Hence, 1440 images are used in this dataset totally.

4) RAF-DB DATASET

The Real-world Affective Face Database (RAF-DB) [12] is a real-world dataset that contains 29,672 highly diverse facial images downloaded from the Internet. With manually crowd-sourced annotation and reliable estimation, seven basic and eleven compound expression labels are provided for the samples. In our experiment, only images with basic expressions (surprise, fear, disgust, happiness, sadness, anger and neutral) are used, including 12,271 images for training and 3,068 images for test.

B. PREPROCESSING

Before training our network, face alignment is conducted to reduce variation in face scale and in-plane rotation on each facial image based on the facial landmarks detected with Supervised Descent Method (SDM) [34]. The detected faces are cropped, resized and converted to 48 × 48 gray-scale images. We ignore extra alignment method in RAF-DB because face images have already been aligned. To avoid over-fitting, two types of data augmentation are adopted: rotate and flip horizontal. Firstly, each preprocessed training image is rotated at angles of $\{-15°, -10°, -5°, 5°, 10°, 15°\}$. Secondly, they are flipped horizontally. Finally, there are 14 derived samples in one image. We employ the same preprocessing for both facial expression datasets and CelebA dataset.

Because of CK+ and Oulu-CASIA do not provide specified training and test sets, we employ the most popular 10-fold validation strategy as in the previous methods [11], [14], [21], [23], [32], [37], [38], [41], [46]. To ensure the generalization of our model, each dataset is split into ten groups without subject overlapping between the groups. For each run, nine groups are used for training and the remaining is used for testing. The results are the average of 10 runs. For the experiments on the RAF-DB dataset, we use their official split for training and test.

C. IMPLEMENTATION DETAILS

The training of WS-LGAN contains two stages. In the first stage, we train the AMG on CelebA. The initial learning rate is set to 0.1, which is decreased by 0.1 after every 20 epochs. After we obtain the well-trained AMG, we freeze it and transfer it to facial expression datasets. In the second stage, we use the frozen AMG to generate attention maps, and train the remaining part of WS-LGAN jointly.

Following previous work, before training on the target expression datasets, we pre-train WS-LGAN on FER2013 dataset [35] and fine-tune WS-LGAN on the target expression datasets. The initial learning rate for pre-train and fine-tuning are set to 0.1, 0.01 respectively. They are divided by 10 at 50% and 75% of the total training epochs. We optimize the model using Stochastic Gradient Descent (SGD)

with a batch size of 100, momentum of 0.9, weight decay of 0.0005 for all stages. In Eq.6, λ_3 are set to 2, 5, 2 for CK+, Oulu-CASIA and RAF-DB respectively, while other parameters are set to 1 empirically.

D. EXPRESSION RECOGNITION RESULTS

To evaluate the performance of WS-LGAN, we compare it with other competitive approaches through two indicators, namely feature (dynamic feature or static feature) and average recognition accuracy.

TABLE 2. Performance comparison on the CK+ dataset [30] in terms of the average recognition accuracy.

Method	Feature	Accuracy
HOG 3D [36]	Dynamic	91.44%
Cov3D [40]	Dynamic	92.30%
3DCNN-DAP [24]	Dynamic	92.40%
DCMA-CNNs [14]	Static	93.46%
STM-ExpLet [37]	Dynamic	94.19%
IL-CNN [11]	Static	94.35%
LOMo [41]	Dynamic	95.10%
IACNN [22]	Static	95.37%
PAT-ResNet-(gender,race) [46]	Static	95.82%
BDBN [16]	Static	96.70%
2B(N+M)Softmax [13]	Static	97.10%
DTAGN [38]	Dynamic	97.25%
DeRL [23]	Static	97.30%
GCNet [39]	Static	97.93%
PHRNN-MSCNN [21]	Dynamic	98.50%
WS-LGAN	Static	**98.06%**

TABLE 3. Confusion matrix of the proposed method evaluated on the CK+ dataset [30]. The ground truth and the predicted labels are given by the first column and the first row, respectively.

	An	Co	Di	Fe	Ha	Sa	Su
An	**97.8**	0	0	0	0	2.2	0
Co	0	**94.4**	0	0	0	5.6	0
Di	0	0	**100**	0	0	0	0
Fe	0	0	0	**100**	0	0	0
Ha	0	0	0	0	**100**	0	0
Sa	8.3	0	0	3.6	0	**88.1**	0
Su	0	1.2	0	0	0	0	**98.8**

1) RESULTS ON CK+ DATASET.

Table 2 shows the results of the comparative studies in terms of average accuracy on CK+. Table 3 is the confusion matrix, which illustrates the detailed classification results of all seven expressions. The diagonal entries represent the recognition accuracy for each expression. From Table 2 and Table 3, we draw the following conclusions.

(1) Our method achieves an average recognition accuracy of 98.06% on CK+. Among the methods which utilize only static image, our result achieves state-of-the-art. It seems that our performance is a little worse than PHRNN-MSCNN [21]. This is because PHRNN-MSCNN utilizes partial-whole, geometry-appearance and dynamic-still features. Motion information is added to their model and their inputs are complex. While our method needs only static appearance features, which is more favorable for online applications or snapshots where per frame labels are preferred.

(2) It performs well on disgust, fear and happiness, but the performances on contempt and sadness are poor. The low accuracy of contempt is mainly due to the lack of data. The samples of contempt are only 18/327 of the total, which is far less than others. Besides, sadness and anger are confused in some samples. A reasonable explanation is that sadness and anger share some similar actions in local facial regions [42]. In Facial Action Coding System (FACS) these two expressions have shared Action Units (AUs): AU4 (Brow Lowerer) and AU17 (Chin Raiser) [43], [44].

(3) WS-LGAN obtains a better recognition accuracy than IACNN [22] and 2B(N+M)Softmax [13], which aim to obtain more discriminative representation for holistic facial image by metric learning. The results show that more attention to local regions will facilitate expression classification.

TABLE 4. Performance comparison on the Oulu-CASIA dataset [33] in terms of the average accuracy.

Method	Feature	Accuracy
HOG 3D [36]	Dynamic	70.63%
STM-ExpLet [37]	Dynamic	74.59%
IL-CNN [11]	Static	77.29%
DTAGN [38]	Dynamic	81.86%
LOMo [41]	Dynamic	82.10%
PHRNN-MSCNN [21]	Dynamic	86.25%
GCNet [39]	Static	86.39%
FN2EN [32]	Static	87.71%
DeRL [23]	Static	88.00%
WS-LGAN	Static	**88.26%**

TABLE 5. Confusion matrix of the proposed method evaluated on the Oulu-CASIA dataset [33]. The ground truth and the predicted labels are given by the first column and the first row, respectively.

	An	Di	Fe	Ha	Sa	Su
An	**77.1**	14.1	4.2	0	4.6	0
Di	6.3	**85.0**	2.4	0	6.3	0
Fe	1.3	0	**84.2**	5.8	2.5	6.2
Ha	1.3	0	2.9	**95.8**	0	0
Sa	5.0	2.9	0	1.3	**90.8**	0
Su	0	0	3.3	0	0	**96.7**

2) RESULTS ON OULU-CASIA DATASET

For Oulu-CASIA dataset, the results are reported in Table 4. Details of classification results are shown in the confusion matrix in Table 5. From Table 4 and Table 5, discussions can be summarized as the following.

(1) Our method achieves an average recognition accuracy of 88.26% on Oulu-CASIA. The performance of WS-LGAN is better than all the state-of-the-art methods, including methods that use dynamic or static features. Note that, Oulu-CASIA is a more challenging dataset, which includes changes in facial attributes, such as with glasses on. WS-LGAN can still correctly classify most expressions, which demonstrates the robustness of the proposed method.

(2) WS-LGAN performs well when recognizing happiness and surprise, which reach the accuracy of 95.8% and 96.7%, respectively. Disgust and anger, sadness and anger are seriously confused. The main reason is they act similarly in some

facial action units in FACS, such as AU10 (Upper Lip Raiser), AU17 (Chin Raiser), AU25 (Lips Part) and AU26 (Jaw Drop) involved in disgust and anger, AU4 (Brow Lowerer) and AU17 (Chin Raiser) involved in sadness and anger [42]–[44].

TABLE 6. Performance comparison on the RAF-DB [12] dataset in terms of the average accuracy. Some papers report performance as an average of diagonal values of confusion matrix. We convert them to regular accuracy for fair comparison.

Method	Feature	Accuracy
FSN [45]	Static	81.10%
baseDCNN [12]	Static	82.86%
Center Loss [12]	Static	83.68%
DLP-CNN [12]	Static	84.13%
PAT-ResNet-(gender,race) [46]	Static	84.19%
Lin et al. [47]	Static	84.68%
WS-LGAN	Static	**85.07%**

TABLE 7. Confusion matrix of the proposed method evaluated on the RAF-DB dataset [12]. The ground truth and the predicted labels are given by the first column and the first row, respectively.

	An	Di	Fe	Ha	Sa	Su	Ne
An	**75.3**	4.3	1.2	8.6	3.7	3.1	3.7
Di	6.9	**54.4**	2.5	8.1	8.1	3.8	16.3
Fe	5.4	1.4	**63.5**	8.1	8.1	8.1	5.4
Ha	0.5	0.8	0.3	**93.8**	2.1	0.3	2.3
Sa	1.7	2.1	1.5	4.2	**83.5**	0.6	6.5
Su	0.6	2.7	3.0	3.0	3.3	**82.7**	4.6
Ne	0.0	2.2	0.3	3.8	8.4	1.3	**84.0**

3) RESULTS ON RAF-DB DATASET

For RAF-DB dataset, the results of comparison are reported in Table 6. Details of classification results are shown in the confusion matrix in Table 7. From Table 6 and Table 7, discussions can be summarized as the following.

(1) The proposed WS-LGAN achieves an average recognition accuracy of 85.07% on RAF-DB which is closer to the natural scene. The performance of WS-LGAN is far better than all methods. It proves that our method is robust to both lab-controlled and real-world facial expression dataset.

(2) The highest accuracy is obtained when recognizing happiness, which reaches to 93.8%. However, the performances on anger, disgust and fear are poor. This is mainly due to the lack of data. In RAF-DB the samples of anger, disgust and fear are far less than others.

E. ABLATION STUDIES

The performance of the network is mainly determined by the following four components: global features, local features, SFU and contrastive loss. To assess these four components, we conduct some ablation experiments on the CK+ dataset to evaluate their effect on recognition.

1) THE EFFECTS OF FEATURE FUSION

We construct another two models to take on the evaluation task. The model only utilizes global features to make classification is denoted as GFNet. The model that recognizes expressions only with local features is denoted as LFNet.

TABLE 8. Recognition accuracy on the CK+ dataset with different types of features.

Model	Accuracy
WS-LGAN	98.06%
GFNet	95.10%
LFNet	94.90%

The recognition performances of these two models are listed in Table 8.

From Table 8, we can observe that the recognition accuracy of WS-LGAN is much higher than GFNet and LFNet, which means that facial expression recognition benefit from feature fusion. This is reasonable as global features or local features only focus on representing expressional information with a specific aspect. The global feature is intended to represent the integrity of the expression, while the local feature focuses on the subtle traits of the local region. The improvement on recognition accuracy by fusion indicates that these two types of features are complementary to each other.

2) THE EFFECTS OF THE SFU

In our model, feature fusion is crucial to the final recognition performance. In addition to the SFU, we also explore the properties of sum fusion and concatenation fusion. Sum fusion computes the sum of all feature maps at the same spatial location and feature channel. Since the channel numbering is arbitrary, sum fusion defines an arbitrary combination between feature channels. The model with sum fusion is denoted as WS-LGAN-Sum. Concatenation fusion stacks the two feature maps at the same spatial location across the feature channels. All features are treated with the same confidence. The model with concatenation fusion is denoted as WS-LGAN-Concat. Experimental results on the CK+ dataset are summarized in Table 9. Our WS-LGAN achieves the highest accuracy by fusing features through the SFU. The excellent performance of the SFU can be attributed to the adaptive weighted mechanism among different features.

TABLE 9. Recognition accuracy on the CK+ dataset with different types of feature fusion strategies.

Model	Accuracy
WS-LGAN	98.06%
WS-LGAN-Sum	95.82%
WS-LGAN-Concat	96.63%

3) THE EFFECTS OF CONTRASTIVE LOSS

In this experiment, the model without contrastive loss is denoted as WS-LGAN-WCL. In other words, in WS-LGAN-WCL, we only use the softmax loss as supervision signal to optimize the parameters. We compare the performance of WS-LGAN-WCL with the proposed model. The recognition result is shown in Table 10.

From Table 10, we can see that the proposed model performs better than WS-LGAN-WCL. This is reasonable as softmax loss forces the features of different expressions staying apart, but it has not a strong constraint to reduce the

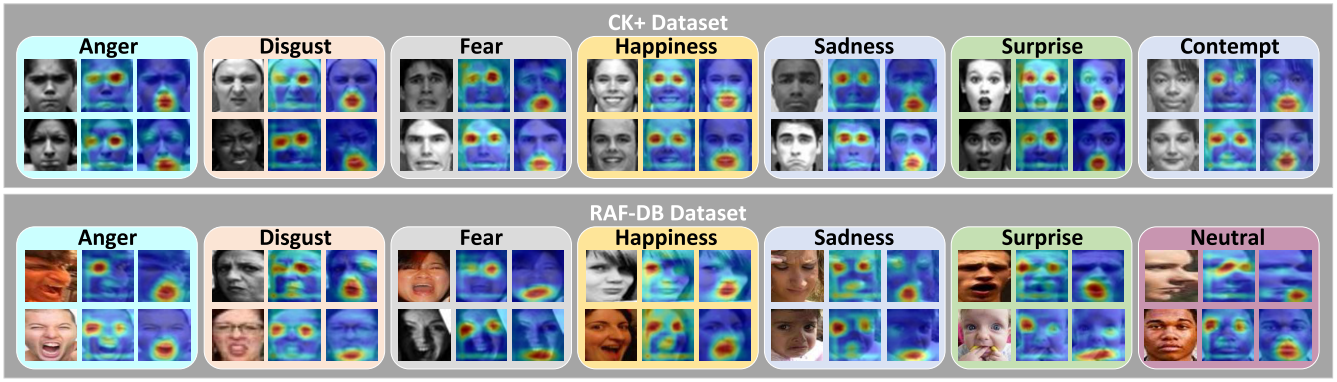


FIGURE 6. Visualization of the attention maps generated on the CK+ dataset and the RAF-DB dataset. Best view in color.

TABLE 10. Effects of contrastive loss on the CK+ dataset.

Model	Accuracy
WS-LGAN	98.06%
WS-LGAN-Sum	95.82%
WS-LGAN-Concat	96.63%

variations of identical expressions. The two local contrastive losses and one global contrastive loss correspond to local representations and global representation work together to push our model to focus on expression details in different granularities. With the joint supervision of softmax loss, local contrastive loss and global contrastive loss, not only the inter-class features differences are enlarged, but also the intra-class features variations are reduced. Hence the discriminative power of the learned features can be highly enhanced. The improvement in recognition accuracy demonstrates the effectiveness of contrastive loss.

F. VISUALIZATION OF ATTENTION MAPS

In Figure 6, we visualize the attention maps generated by transfer the AMG to CK+ and RAF-DB datasets to demonstrate the effectiveness of weakly supervised attention learning. Rectangular boxes of different colors contain visualized results of different expressions. Within each rectangular box, the first column is the original images, the second column is the eye-related attention maps, and the last column is the mouth-related attention maps. We can see that, regardless of the person or expression in the image, our model can always accurately locate the eyes region and mouth region. It provides an efficient and accurate guidance for the extraction of local features. In addition, it avoids the introduction of many unrelated factors compared to using all face patches.

V. CONCLUSION

In this paper, we propose a Weakly Supervised Local-Global Attention Network to perform facial expression recognition with jointly use of local and global features. Our approach shows how we can directly locate the crucial regions and extract the corresponding local features under weak supervision. Selective Feature Unit is designed to fusion local and global features in an adaptive manner. It enables two types of features to complement each other to boost the recognition performance. Besides, contrastive loss is introduced for both local and global features to increase inter-class differences and decrease intra-class variations under different granularities. Experimental results on three databases demonstrate that our proposed methods have achieved the state-of-the-art performance.

Furthermore, the approach of perceiving crucial local regions proposed in this work has potential application value for other face related tasks, such as face detection, face alignment and face attribute manipulation.

REFERENCES

[1] E. Vural, M. Cetin, A. Ercil, G. Littlewort, M. Bartlett, and J. Movellan, "Automated drowsiness detection for improved driving safety," in *Proc. Int. Conf. Automat. Technol.*, 2008, pp. 96–105.
[2] A. Ryan, J. F. Cohn, S. Lucey, J. Saragih, P. Lucey, F. De la Torre, and A. Rossi, "Automated facial expression recognition system," in *Proc. 43rd Annu. Int. Carnahan Conf. Secur. Technol.*, Oct. 2009, pp. 172–177.
[3] K. Sikka, A. Dhall, and M. S. Bartlett, "Classification and weakly supervised pain localization using multiple segment representation," *Image Vis. Comput.*, vol. 32, no. 10, pp. 659–670, Oct. 2014.
[4] P. Lewinski, M. L. Fransen, and E. S. H. Tan, "Predicting advertising effectiveness by facial expressions in response to amusing persuasive stimuli," *J. Neurosci., Psychol., Econ.*, vol. 7, no. 1, pp. 1–14, 2014.
[5] N. T. Cao, A. H. T. That, and H. I. Choi, "An effective facial expression recognition approach for intelligent game systems," *Int. J. Comput. Vis. Robot.*, vol. 6, no. 3, p. 223, 2016.
[6] 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.
[7] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 6, pp. 1113–1133, Jun. 2015.
[8] S. Li and W. Deng, "Deep facial expression recognition: A survey," 2018, *arXiv:1804.08348*. [Online]. Available: https://arxiv.org/abs/1804.08348
[9] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 499–515.
[10] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.
[11] J. Cai, Z. Meng, A. S. Khan, Z. Li, and Y. Tong, "Island loss for learning discriminative features in facial expression recognition," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, May 2018 pp. 302–309.
[12] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2852–2861.
[13] X. Liu, B. V. K. 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. 522–531.

[14] 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.

[15] 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.

[16] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1805–1812.

[17] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Rev. Neurosci.*, vol. 2, no. 3, pp. 194–203, Mar. 2001.

[18] J. D. Boucher and P. Ekman, "Facial areas and emotional information," *J. Commun.*, vol. 25, no. 2, pp. 21–29, Feb. 2006.

[19] J. F. Cohn and A. Zlochower, "A computerized analysis of facial expression: Feasibility of automated discrimination," *Amer. Psychol. Soc.*, vol. 2, p. 6, 1995.

[20] 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.

[21] K. Zhang, Y. Huang, Y. Du, and L. Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4193–4203, Sep. 2017.

[22] Z. Meng, P. Liu, J. Cai, S. Han, and Y. Tong, "Identity-aware convolutional neural network for facial expression recognition," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2017, pp. 558–565.

[23] 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. 2220–2224.

[24] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen, "Deeply learning deformable facial action parts model for dynamic expression analysis," in *Proc. IEEE Asian Conf. Comput. Vis.*, Nov. 2014, pp. 143–157.

[25] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[26] 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.

[27] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2–7.

[28] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear convolutional neural networks for fine-grained visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1309–1322, Jun. 2018.

[29] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 510–519.

[30] 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. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.-Workshops*, Jun. 2010, pp. 94–101.

[31] T. Kanade, J. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proc. 4th IEEE Int. Conf. FGR*, Mar. 2000, pp. 46–53.

[32] 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.

[33] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen, "Facial expression recognition from near-infrared videos," *Image Vis. Comput.*, vol. 29, no. 9, pp. 607–619, Aug. 2011.

[34] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 532–539.

[35] 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.

[36] A. Klaeser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *Proc. Brit. Mach. Vis. Conf.*, 2008, pp. 1–275.

[37] M. Liu, S. Shan, R. Wang, and X. Chen, "Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1749–1756.

[38] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2983–2991.

[39] Y. Kim, B. Yoo, Y. Kwak, C. Choi, and J. Kim, "Deep generative-contrastive networks for facial expression recognition," 2017, *arXiv:1703.07140*. [Online]. Available: https://arxiv.org/abs/1703.07140

[40] A. Sanin, C. Sanderson, M. T. Harandi, and B. C. Lovell, "Spatio-temporal covariance descriptors for action and gesture recognition," in *Proc. IEEE Workshop Appl. Comput. Vis. (WACV)*, Jan. 2013, pp. 103–110.

[41] K. Sikka, G. Sharma, and M. Bartlett, "LOMo: Latent ordinal model for facial analysis in videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5580–5589.

[42] B. Martinez, M. F. Valstar, B. Jiang, and M. Pantic, "Automatic analysis of facial actions: A survey," *IEEE Trans. Affect. Comput.*, vol. 10, no. 3, pp. 325–347, Jul. 2019.

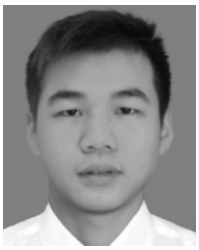
[43] W. V. Friesen and P. Ekman, "EMFACS-7: Emotional facial action coding system," *Unpublished Manuscript, Univ. California San Francisco*, vol. 2, no. 36, p. 1, 1983.

[44] P. Ekman and W. V. Friesen, *Facial Action Coding System (FACS): Manual*. Palo Alto, CA, USA: Consulting Psychologist Press, 1978.

[45] S. Zhao, H. Cai, H. Liu, J. Zhang, and S. Chen, "Feature selection mechanism in CNNs for facial expression recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2018, p. 317.

[46] J. Cai, Z. Meng, A. Shehab Khan, Z. Li, J. O'Reilly, and Y. Tong, "Probabilistic attribute tree in convolutional neural networks for facial expression recognition," 2018, *arXiv:1812.07067*. [Online]. Available: http://arxiv.org/abs/1812.07067

[47] F. Lin, R. Hong, W. Zhou, and H. Li, "Facial expression recognition with data augmentation and compact feature learning," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 1957–1961.



HAIFENG ZHANG was born in 1993. He received the B.E. degree from the China University of Geosciences, Beijing, in 2015. He is currently pursuing the Ph.D. degree with the University of Science and Technology of China. His work published in several different journals and conferences. His research interests are in face recognition and facial expression recognition.

WEN SU was born in 1992. She received the B.E. degree in engineering from the Automation Department, University of Science and Technology of China, in 2013, and the Ph.D. degree in control science and engineering from the University of Science and Technology of China, in 2018. Her work has been published in several different journals and conferences. She is currently a Lecturer at Zhejiang Sci-Tech University. Her research interests are in semantic segmentation based on deep learning, monocular depth estimating, and 3D scene understanding.

ZENGFU WANG (Member, IEEE) received the B.S. degree in electronic engineering from the University of Science and Technology of China, in 1982, and the Ph.D. degree in control engineering from Osaka University, Japan, in 1992. He is currently a Professor with the Institute of Intelligent Machines, Chinese Academy of Sciences, and the University of Science and Technology of China. He has published more than 200 journal articles and conference papers. His research interests include computer vision, human–computer interaction, and intelligent robots. He received the Best Paper Award at ACM International Conference on Multimedia (ACM Multimedia), in 2009.