Weakly Supervised Local-Global Relation Network for Facial Expression Recognition

Haifeng Zhang[1] , **Wen Su**[3] , **Jun Yu**[1] and **Zengfu Wang**[1,2*]
[1]Department of Automation, University of Science and Technology of China
[2]Institute of Intelligent Machines, Chinese Academy of Sciences
[3]Faculty of Mechanical Engineering and Automation, Zhejiang Sci-Tech University
hfz@mail.ustc.edu.cn, wensu@zstu.edu.cn, {harryjun, zfwang}@ustc.edu.cn

Abstract

To extract crucial local features and enhance the complementary relation between local and global features, this paper proposes a Weakly Supervised Local-Global Relation Network (WS-LGRN), which uses the attention mechanism to deal with part location and feature fusion problems. Firstly, the Attention Map Generator quickly finds the local regions-of-interest under the supervision of image-level labels. Secondly, bilinear attention pooling is employed to generate and refine local features. Thirdly, Relational Reasoning Unit is designed to model the relation among all features before making classification. The weighted fusion mechanism in the Relational Reasoning Unit makes the model benefit from the complementary advantages between different features. In addition, contrastive losses are introduced for local and global features to increase the inter-class dispersion and intra-class compactness at different granularities. Experiments on lab-controlled and real-world facial expression dataset show that WS-LGRN achieves state-of-the-art performance, which demonstrates its superiority in FER.

1 Introduction

Driven by recent advances in human-centered computing, recognizing expressions from facial images has been a popular problem in the field of computer vision, and many studies have been conducted. It can be divided into two categories. One category focuses on learning global representation, while another pays more attention to extract partial discriminative features.

For the first category, a popular approach is to enhance the discriminative power of the deeply learned features by proposing novel loss layers to replace or assist the supervision of the softmax loss [Cai *et al.*, 2018b; Li and Deng, 2018]. 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.*, 2017; Liu *et al.*, 2017; Ali and Hughes, 2019; Yang *et al.*, 2018]. It aims to alleviate variations introduced by identity and achieve identity-invariant FER. However, these methods mentioned above usually extract features from the holistic facial image and ignore fine-grained information in local facial regions. For the second category, the basic premise of learning discriminative part features is that the parts should be located. Some part-based methods crop facial expression images into patches and try to learn local representations from them [Xie and Hu, 2018; Happy and Routray, 2014; Liu *et al.*, 2014]. Although the obtained results are encouraging, there are still some restrictions. Firstly, dividing image into patches can be time-consuming and computationally expensive. Secondly, manually defined patches may not be optimal. Some patches may have no or even negative impact on FER. In addition, if we only focus on local features, we may lose some supplementary information. Attributes provided by the holistic facial image can also affect expressions significantly.

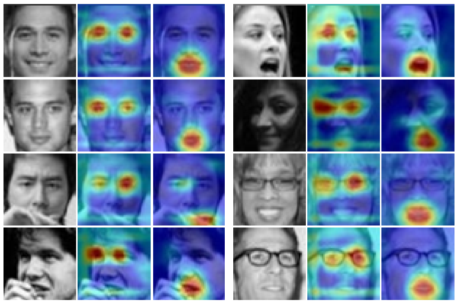


Figure 1: Attention maps that indicates crucial facial regions.

In fact, the human visual attention mechanism shows that humans will first obtain a global description when performing object recognition, and then attention will quickly shift to regions with obvious features [Itti and Koch, 2001]. Besides, results in [Cohn and Zlochower, 1995] indicate that much of expressional clues come from the salient facial regions such as neighbourhood of mouth and eyes. Motivated by these, we propose a Weakly Supervised Local-Global Relation Network (WS-LGRN). Unlike previous methods, we mimic the way humans recognize facial expressions. Specifically, the attention mechanism is introduced to guide our network to locate crucial local regions autonomously and extract

[*]Corresponding Author

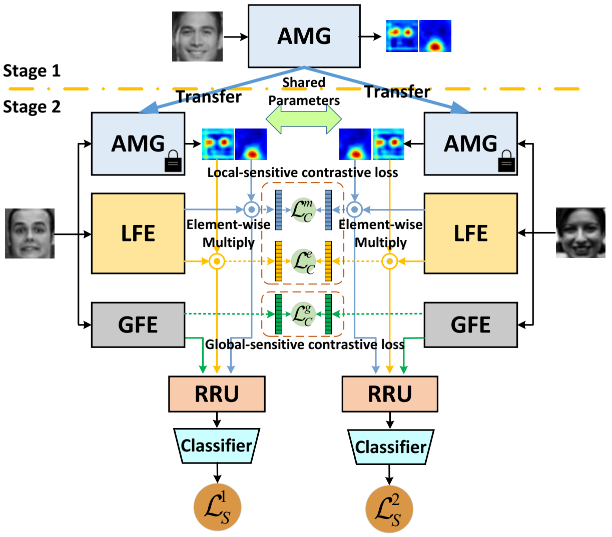


Figure 2: Overview of the proposed framework.

local features through these regions. Since facial expression datasets do not have labeled part locations, we formulate part localization in a weakly supervised manner by introducing a facial attributes dataset. Moreover, we model the relation between local and global features to jointly utilize their complementary advantages to deal with the loss of local details and emphasize global context cues.

During training, our pipeline is decomposed into two stages, as shown in Figure 2. In the first stage, the Attention Map Generator (AMG) is trained on the facial attribute dataset to generate attention maps that designate the regions around eyes and mouth. Figure 1 shows some samples generated by AMG. 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. In the second stage, the well-trained AMG is transferred to facial expression datasets with weights fixed. Therefore, the lack of part annotations in facial expression dataset is well solved. The second stage consists of two identical CNN streams whose weights are shared. It takes a pair of facial expression images as input. In addition to AMG, each CNN stream contains four sub-parts: Local Feature Extractor (LFE), Global Feature Extractor (GFE), Relational Reasoning Unit (RRU) and Classifier. LFE extracts features from holistic facial image. Based on the outputs of AMG and LFE, the local features are extracted and refined by bilinear attention pooling. GFE extracts global features directly from holistic image. RRU aims to fuse all features and model the complementary relation among them. A softmax classifier is used for the final expression classification. We optimize the parameters by simultaneously minimizing the softmax loss, 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 the hybrid features.

To sum up, our main contributions are as follows. (1) Unlike local-based methods that rely on facial patches [Xie and Hu, 2018; Happy and Routray, 2014; Liu *et al.*, 2014], we propose to deal with local features by directly locating crucial regions and extracting 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 the attention maps, a bilinear attention pooling is proposed to generate and refine local features. Besides, weak supervision allows us to overcome the limitation of no part annotations in facial expression dataset. (2) Different from [Xie and Hu, 2018] which fuses local and global features through concatenate fusion, we formulate a RRU to model the complementary relation among all features. The adaptive weight in RRU makes a reasonable trade-off and selection of all features as well as makes the model can benefit from local-global complementary advantages. (3) We extend metric learning to both local and global features to increase inter-class differences as well as reduce intra-class variations at different granularities. Previous methods only employ similarity metrics on the global representation [Meng *et al.*, 2017; Cai *et al.*, 2018b; Li and Deng, 2018; Liu *et al.*, 2017], and fine-grained features are not well learned. In our method, explicit local features make it possible to employ local similarity metric. (4) To demonstrate the superiority of our proposed method, we employ experiments on lab-controlled facial expression datasets (CK+) and real-world facial expression dataset (RAF-DB). Our facial expression recognition solution achieves state-of-the-art results on CK+ and RAF-DB with accuracies of 98.37% and 85.20%, respectively.

2 Proposed Method

2.1 Attention Map Generator

A direct method for locating crucial facial regions is to use image and its pixel-wise segmentation as input and target respectively. However, it requires label maps with pixel-wise annotations, which are expensive to collect. 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.*, 2016] enable the classification network to have remarkable localization ability despite being trained on only image-level labels. Inspired by them, we use attention map to locate crucial facial regions. The attention map is a weight map, which highlights the positions of the crucial regions by giving them higher values. To generate the attention maps, we designed our AGM.

Facial expression datasets usually have only expression labels, while the image in the CelebA dataset [Liu *et al.*, 2015] is labeled with 40 facial attributes. Some attributes can guide AMG training to locate crucial regions. Since we only focus on regions related to facial expressions, we choose facial attributes related to eyes and mouth and divide them into two groups according to their respective facial parts. The grouped attributes are summarized 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.

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

Table 1: Facial attributes grouping.

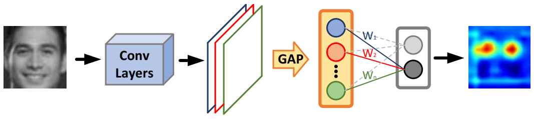


Figure 3: One branch in the Attention Map Generation.

For the training of mouth-related branches, we use the same configuration. Note that, we only use the CelebA dataset to train AMG.

AMG consists of two branches with the same structure for locating eyes and mouth, respectively. Figure 3 shows the branch for eyes. If the image dose not contain any eyes-related attributes listed in Table 1, we take it as a negative example, otherwise it will be a positive example. We use dataset containing these positive and negative examples to train the eyes-related branch. As illustrated in Figure 3, global average pooling (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 features for classification. We back the weights of the output layer to the convolutional features and calculate the weighted sum of the feature maps to obtain our attention maps. We normalize the attention maps so that all values fall in the range [0, 1]. Figure 1 illustrates the effect of attention maps outputted using 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, AMG is frozen.

2.2 Local Feature Refinement

Bilinear Attention Pooling. Firstly, well-trained AMG with fixed weights is used to generate attention maps $A_e \in 1^{1\times H\times W}$ (eyes-related attention maps) and $A_m \in \mathrm{R}^{1\times H\times W}$ (mouth-related attention maps) respectively. Then, we element-wise multiplies feature maps $F \in \mathrm{R}^{C\times H\times W}$ by attention maps A_e and A_m, as shown in Eq.1:

$$F_e = A_e \odot F, \qquad F_m = A_m \odot F. \tag{1}$$

Feature maps F are extracted by LFE from the holistic image. F_e and F_m reflect the feature maps of eyes and mouth, respectively. An example of refining eyes features is shown in Figure 4.

Bilinear attention pooling explicitly define two streams to locate and extract features respectively. We regard the AMG branch as the dorsal stream that deals with the spatial location of the object in the human visual cortex and the LFE branch as the ventral stream that performs object recognition in the human visual cortex. The bilinear attention pooling bridges the appearance models and part locating models. It provides a solution for local feature extraction.

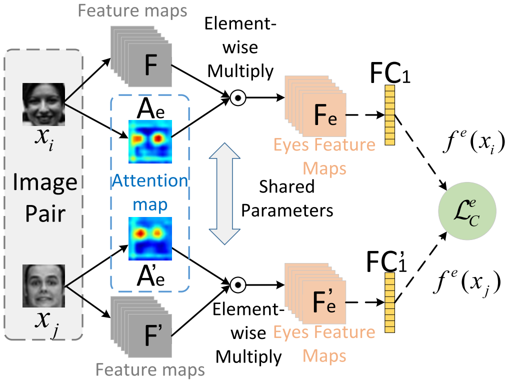


Figure 4: The process of refining eyes features.

Local-Sensitive Contrastive Loss. In order to reduce the intra-class variations and increase the inter-class differences at a 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 draws the eye-related features extracted from samples of the same expression closer to each other, while pushing the eye-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:

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

where x_i and x_j are a pair of training images, and $f^e(x_i)$ and $f^e(x_j)$ are their eyes-related feature vectors. $\theta_{ij} = 1$ means that x_i and x_j are belong to the same facial expression. While $\theta_{ij} = 0$, it reverses. 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 for mouth-related features is defined similar to L_C^e.

2.3 Local-Global Fusion

Global-Sensitive Contrastive Loss. Global feature maps $F_g \in \mathrm{R}^{C\times H\times W}$ are extracted by GFE from holistic facial image directly. A global-sensitive contrastive loss L_C^g is designed for global feature to reduce the intra-class variations and enlarge the inter-class differences. The global feature vector used to calculate the loss function is obtained by inputting F_g into the FC layer. L_C^g is defined as follows:

$$\begin{aligned}&L_C^g(\theta_{ij}, f^g(x_i), f^g(x_j))\\ &= \begin{cases} \frac{1}{2}(\left\|f^g(x_i) - f^g(x_j)\right\|_2^2 & if\theta_{ij} = 1\\ \frac{1}{2}\max\left(0, \delta^g - \left\|f^g(x_i) - f^g(x_j)\right\|_2\right)^2 & if\theta_{ij} = 0\end{cases}\end{aligned} \tag{3}$$

where $f^g(x_i)$ and $f^g(x_j)$ are global feature vectors for a pair of training samples. $\theta_{ij} = 1$ means that x_i and x_j are

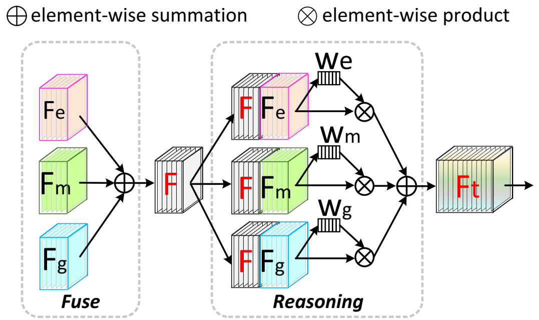


Figure 5: Relational Reasoning Unit.

belong to the same facial expression. While $\theta_{ij} = 0$, it reverses. δ^g is the size of the margin which determines how much dissimilar pairs contribute to the loss function. It is set to 10 empirically.

Relational Reasoning Unit. RRU is designed to model the complementary relation among eyes features, mouth features and global features. Specifically, RRU consists of two key operators: fuse and reasoning as illustrated in Figure 5.

Fuse: To model the complementary relation among 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 next layer. The gates integrate information from all branches. We obtain the hybrid representation from three branches via an element-wise summation:

$$F = F_e + F_m + F_g \tag{4}$$

F is used as the anchor of relational reasoning for learning content-aware attention weight.

Reasoning: The reasoning operator is a attention mechanism on the concatenation of individual feature and hybrid representation for relational reasoning. The design philosophy behind reasoning is to constrain the complementary relation among all features so that it captures the content-aware attention weight of relational reasoning. The weight makes a reasonable trade-off and selection of all features. Specifically, we use concatenation and FC layer to adaptively compute the attention weight for three different spatial descriptors: F_e, F_m and F_g. In its simplest form the weight calculation is a composite function:

$$w_e = g(f_\varphi([F : F_e])) \tag{5}$$

$$w_m = g(f_\varphi([F : F_m])) \tag{6}$$

$$w_g = g(f_\varphi([F : F_g])) \tag{7}$$

For our purposes g and f_φ are sigmoid function and FC, respectively. φ is the parameter of FC. We can call the learned weight a ”relation”; therefore, the role of w_e, w_m, w_g are to infer the ways in which two features are related, or if they are even related at all. Finally, we aggregate all the individual feature along with the hybrid representation into a new compact feature as,

$$F_t = w_e[F : F_e] + w_m[F : F_m] + w_g[F : F_g] \tag{8}$$

F_t is used as the final representation of the proposed RRU for the classification. After RRU, F_t is fed into the Classifier.

2.4 Total Loss

Softmax loss that calculates the classification errors is used on end of each CNN stream to ensure the learned features are meaningful for FER. Combining the two local-sensitive contrastive losses and one global-sensitive contrastive loss mentioned above, the total loss of WS-LGRN is:

$$L_{total} = \lambda_1 L_C^e + \lambda_2 L_C^m + \lambda_3 L_C^g + \lambda_4 L_S^1 + \lambda_5 L_S^2 \tag{9}$$

where $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6\}$ are the weights of each loss. L_S^1 and L_S^2 are the final classification errors.

3 Experiments

3.1 Dataset and Preprocessing

Most of our experiments are conducted on the CK+ [Lucey *et al.*, 2010] dataset. It is a lab-controlled dataset which is annotated with seven expressions, i.e. Anger (An), Disgust (Di), Fear (Fe), Happiness (Ha), Sadness (Sa), Surprise (Su) and Contempt (Co). It consists of 327 facial expression sequences collected from 118 different subjects. Each sequence starts with a neutral expression and ends with a peak expression. As a general procedure [Cai *et al.*, 2018b; Meng *et al.*, 2017; Ali and Hughes, 2019; Ding *et al.*, 2017; Chen *et al.*, 2019], the last three frames of each sequence are used for training and test. Thus, CK+ contains 981 images for our experiments. Additionally, we also conduct experiments on the Real-world Affective Face Database (RAF-DB) [Li and Deng, 2018]. It is a real-world dataset that contains 29,672 highly diverse facial images downloaded from the Internet. Images with seven basic expressions (surprise, fear, disgust, happiness, sadness, anger and neutral) are used in our experiment, including 12,271 images for training and 3,068 images for test.

Face alignment is conducted based on the facial landmarks detected with Supervised Descent Method (SDM) [Xiong and La Torre, 2013]. The detected face are cropped, resized and converted to 48 × 48 grayscale 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. First, each preprocessed training image is rotated at angles of $\{-15°, -10°, -5°, 0°, 5°, 10°, 15°\}$. Then, they are flipped horizontally. We employ same preprocessing for both facial expression datasets and CelebA dataset. Because CK+ does not provide specified training and test sets, we employ the most popular 10-fold validation strategy as in the previous methods [Ali and Hughes, 2019; Ding *et al.*, 2017; Cai *et al.*, 2018b; Chen *et al.*, 2019]. The 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 test. The results are the average of 10 runs. For the experiments on the RAF-DB database, we use their official split for training and test.

3.2 Implement Details

The backbone of each branch in the AMG is a variant of Densenet. It consists of 3 dense block 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 LFE and GFE after reduce the number of dense layers to 6 for each dense block. All of the pooling layers in the transition layer are 2 × 2 average pooling with stride 2. The training of WS-LGRN contains two stages. In the first stage, we train 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-LGRN jointly. Following previous works, before training on the target expression datasets, we pre-train WS-LGRN on FER2013 dataset [Goodfellow *et al.*, 2015] and fine-tune WS-LGRN 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 with a batch size of 100, momentum of 0.9, weight decay of 0.0005 for all stages. In Eq.9, λ_1 is set to 3 for CK+ and RAF-DB, while other parameters are set to 1 empirically.

3.3 Ablation Studies

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

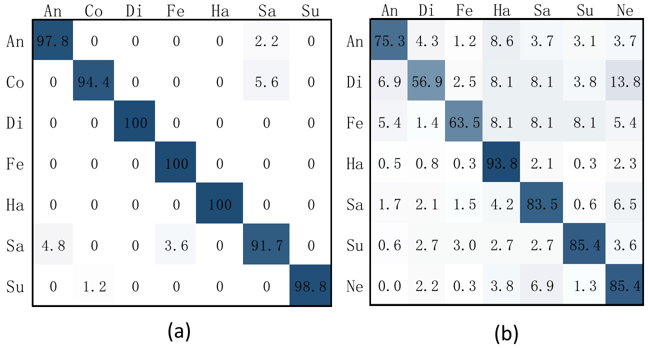
The effects of feature fusion. The model only utilizes global features to make classification is denoted as GFNet. The model that recognizes expressions only with local features is denoted as LFNet. From Table 2, we can observe that the recognition accuracy of WS-LGRN is much higher than GFNet and LFNet, which means FER benefits 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.

The effects of the RRU. In our model, RRU fuses all features and considers their complementary relation. In addition to the RRU, 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. The model with sum fusion is denoted as WS-LGRN-Sum. Concatenation fusion stacks the two feature maps at the same spatial location across the feature channels. The model with concatenation fusion is denoted as WS-LGRN-Concat. Experimental results are summarized in Table 2. Our WS-LGRN achieves the highest accuracy by fusing features through the RRU. RRU can adaptively capture the importance of each individual feature, and make a reasonable trade-off between local and global features.

The effects of contrastive loss. In this experiment, the model which only uses the softmax loss to optimize the parameters is denoted as WS-LGRN-WCL. We compare the performance of WS-LGRN-WCL with the proposed model.

Model	Accuracy(%)
WS-LGRN	98.37
GFNet	95.10
LFNet	94.90
WS-LGRN-Sum	96.13
WS-LGRN-Concat	96.94
WS-LGRN-WCL	97.35

Table 2: Recognition accuracy on the CK+ dataset with different types of features.



(a)

	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	4.8	0	0	3.6	0	91.7	0
Su	0	1.2	0	0	0	0	98.8

(b)

	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	56.9	2.5	8.1	8.1	3.8	13.8
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	2.7	2.7	85.4	3.6
Ne	0.0	2.2	0.3	3.8	6.9	1.3	85.4

Figure 6: Confusion matrices on the CK+ (a) and RAF-DB (b).

From Table 2, we can see that the proposed model performs better than WS-LGRN-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 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. The improvement in recognition accuracy demonstrates the effectiveness of contrastive loss.

With the simultaneous use of global features, local features, RRU and contrastive losses, we obtained the best recognition performance. Therefore, we will use the same configuration in the following experiments.

3.4 Expression Recognition Results

To evaluate the overall performance, the confusion matrices on two datasets are illustrated in Figure 6. To compare the performance of the proposed method with other methods, Table 3 and Table 4 list the accuracy of our proposed and the state-of-the-art methods on the CK+ and RAF-DB databases.

Results on CK+ dataset. Our method achieves an average recognition accuracy of 98.37% on CK+. Among the methods which utilize only static image, our result achieves state-of-the-art. Our method 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

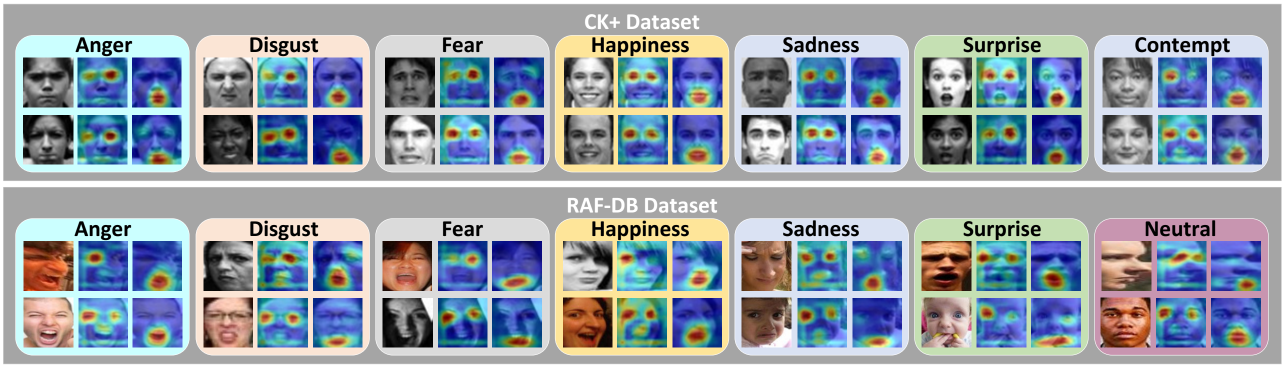


Figure 7: Visualization of the attention maps generated on the CK+ and RAF-DB dataset. Best view in color.

Method	Accuracy(%)
IL-CNN [Cai *et al.*, 2018b]	94.35
IACNN [Meng *et al.*, 2017]	95.37
PAT-ResNet-(gender,race) [Cai *et al.*, 2018a]	95.82
2B(N+M)Softmax [Liu *et al.*, 2017]	97.10
DE-GAN [Ali and Hughes, 2019]	97.28
DeRL [Yang *et al.*, 2018]	97.30
FMPN [Chen *et al.*, 2019]	98.06
WS-LGRN	98.37

Table 3: Performance comparison on the CK+ dataset.

Method	Accuracy(%)
FSN [Zhao *et al.*, 2018]	81.10
baseDCNN [Li and Deng, 2018]	82.86
Center Loss [Li and Deng, 2018]	83.68
DLP-CNN [Li and Deng, 2018]	84.13
PAT-ResNet-(gender,race) [Cai *et al.*, 2018a]	84.19
Lin et al. [Lin *et al.*, 2018]	84.68
APM-VGG [Li *et al.*, 2019]	85.17
WS-LGRN	85.79

Table 4: Performance comparison on the RAF-DB dataset.

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.

Results on RAF-DB dataset. Our method achieves an average recognition accuracy of 85.20% on RAF-DB which is a dataset closer to the natural scene. It is better than all methods. Notice that, some papers report performance as an average of diagonal values of confusion matrix. We convert them to regular accuracy for fair comparison. It proves that our method is robust to both lab-controlled and real-world facial expression dataset. The highest accuracy is obtained when recognizing happiness, which reaches to 93.8%. However, the performance 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.

3.5 Visualization of Attention Maps

In Figure 7, we visualize the attention maps generated by transfer AMG to CK+ and RAF-DB 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 picture, our model can always accurately locate the eye region and mouth region. This provides an efficient and accurate guidance for the extraction of local features. In addition, this avoids the introduction of many unrelated factors compared to using all face patches.

4 Conclusions

In this paper, we proposed a weakly supervised local attention network which automatically perceives the crucial local regions of the face, so that the network can focus on representative local features while acquiring the global facial features. In the proposed WS-LGRN, an Attention Map Generator trained on facial attributes dataset under weakly supervision is adopted to perceive the location of crucial local regions. Local feature refinement is employed by bilinear attention pooling. Contrastive loss is introduced for both local and global features to increase inter-class differences and decrease intra-class variations under different scale. Relation Reasoning Unit is designed to model the complementary relation of local and global features. Extensive experiments on lab-controlled and real-world datasets demonstrate the effectiveness of our proposed method.

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.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No.61472393).

References

[Ali and Hughes, 2019] Kamran Ali and Charles E Hughes. Facial expression recognition using disentangled adversarial learning. *arXiv preprint arXiv:1909.13135*, 2019.

[Cai *et al.*, 2018a] Jie Cai, Zibo Meng, Ahmed Shehab Khan, Zhiyuan Li, James O'Reilly, and Yan Tong. Probabilistic attribute tree in convolutional neural networks for facial expression recognition. *arXiv preprint arXiv:1812.07067*, 2018.

[Cai *et al.*, 2018b] Jie Cai, Zibo Meng, Ahmed Shehab Khan, Zhiyuan Li, James O'Reilly, and Yan Tong. Island loss for learning discriminative features in facial expression recognition. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 302–309. IEEE, 2018.

[Chen *et al.*, 2019] Yuedong Chen, Jianfeng Wang, Shikai Chen, Zhongchao Shi, and Jianfei Cai. Facial motion prior networks for facial expression recognition. *arXiv preprint arXiv:1902.08788*, 2019.

[Cohn and Zlochower, 1995] JF Cohn and A Zlochower. A computerized analysis of facial expression: Feasibility of automated discrimination. *American Psychological Society*, 2:6, 1995.

[Ding *et al.*, 2017] Hui Ding, Shaohua Kevin Zhou, and Rama Chellappa. Facenet2expnet: Regularizing a deep face recognition net for expression recognition. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 118–126. IEEE, 2017.

[Goodfellow *et al.*, 2015] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. *Neural Networks*, 64:59–63, 2015.

[Happy and Routray, 2014] SL Happy and Aurobinda Routray. Automatic facial expression recognition using features of salient facial patches. *IEEE transactions on Affective Computing*, 6(1):1–12, 2014.

[Itti and Koch, 2001] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194, 2001.

[Li and Deng, 2018] Shan Li and Weihong Deng. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing*, 28(1):356–370, 2018.

[Li *et al.*, 2019] Zhiyuan Li, Shizhong Han, Ahmed Shehab Khan, Jie Cai, Zibo Meng, James O'Reilly, and Yan Tong. Pooling map adaptation in convolutional neural network for facial expression recognition. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1108–1113. IEEE, 2019.

[Lin *et al.*, 2018] Feng Lin, Richang Hong, Wengang Zhou, and Houqiang Li. Facial expression recognition with data augmentation and compact feature learning. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 1957–1961. IEEE, 2018.

[Liu *et al.*, 2014] Ping Liu, Shizhong Han, Zibo Meng, and Yan Tong. Facial expression recognition via a boosted deep belief network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1805–1812, 2014.

[Liu *et al.*, 2015] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[Liu *et al.*, 2017] Xiaofeng Liu, BVK Vijaya Kumar, Jane You, and Ping Jia. Adaptive deep metric learning for identity-aware facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 20–29, 2017.

[Lucey *et al.*, 2010] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 94–101. IEEE, 2010.

[Meng *et al.*, 2017] Zibo Meng, Ping Liu, Jie Cai, Shizhong Han, and Yan Tong. Identity-aware convolutional neural network for facial expression recognition. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 558–565. IEEE, 2017.

[Xie and Hu, 2018] Siyue Xie and Haifeng Hu. Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks. *IEEE Transactions on Multimedia*, 21(1):211–220, 2018.

[Xiong and La Torre, 2013] Xuehan Xiong and Fernando De La Torre. Supervised descent method and its applications to face alignment. 2013:532–539, 2013.

[Yang *et al.*, 2018] Huiyuan Yang, Umur Ciftci, and Lijun Yin. Facial expression recognition by de-expression residue learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2168–2177, 2018.

[Zhao *et al.*, 2018] Shuwen Zhao, Haibin Cai, Honghai Liu, Jianhua Zhang, and Shengyong Chen. Feature selection mechanism in cnns for facial expression recognition. In *BMVC*, page 317, 2018.

[Zhou *et al.*, 2016] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.