IVSA:Facial Expression Recognition Method with Salient Attention

Shuai Huang[1,3], Dingkang Yang[1,2], Chuyi Zhong[1,3], Lihua Zhang[1,2,3,4,*]
[1]Academy for Engineering and Technology, Fudan University
[2]Ji Hua Laboratory, Foshan, China
[3]Engineering Research Center of AI and Robotics, Shanghai, China
[4]Jilin Provincial Key Laboratory of Intelligence Science and Engineering, Changchun, China
{shuaihuang20,dkyang20,cyzhong20,scyan20,lihuazhang}@fudan.edu.cn

***Abstract*—In recent years, facial expressions have become a hot topic in computer vision research. With the development of artificial intelligence technology, the performance of facial expression recognition have been greatly improved. To further improve the ability of extracting significant features and enhancing the robustness of the model, we present a novel facial expression recognition method based on convolutional neural network and attention mechanism. Concretely, we add the L2 norm features in CBAM and re-scale the channel weights. Salient attention block is used to suppress the insignificant features and enhance the weight of salient features, which improves the performance and robustness of the model. Finally, IVSA achieves the highest single-network accuracy of 72.44% on testing data without using extra training data, which is improved by 1.14% compared with the previous methods. Extensive experiments prove the effectiveness of the model and framework proposed.**

Keywords—deep learning, computer vision, facial expression recognition, convolutional neural network, attention mechanism

I. INTRODUCTION

Facial expression is the most natural and common signal to convey emotional state and intention in people's daily life [1-2]. In recent years, with the rapid development of machine learning and deep learning, facial expression recognition technology has been applied to many fields, such as human-computer interaction [3], intelligent control [1], and intelligent medicine [4]. Facial Expression Recognition (FER) has become an important research subject of computer vision. It plays an important role in practical applications, such as vehicle camera analyzing the driver's physical state through facial expression [5], robot using facial expression to analyze the psychological state of children or the elderly [4, 6], etc.

The methods of facial expression recognition are mainly divided into two categories: manual features-based methods and deep learning-based methods. The manual features include Histogram of Oriented Gradient (HOG) [7], Local Binary Patterns (LBP) [8], Non-negative Matrix Factorization (NMF) [9], etc. These feature extraction methods need to manually specify the extraction features. They will lose some expression feature information, and make it more difficult to extract the high-order features of facial expression. The methods based on deep learning is to automatically extract features from models, such as VGG [10] and ResNet [11], which has better accuracy and greatly exceeds the results obtained by traditional methods.

Although the existing research on facial expression recognition has made advanced progress [12-14], most of them achieve better accuracy by increasing the depth and width of the model, or by combining different models. They do not explore how to improve the performance of a single model without using extra data. In this work, we aim to improve the performance and robustness of the single-network model by integrating convolutional neural network (CNN) and attention. We first adopt the CNN network and then integrate the attention module to the improved CNN, which aims to suppress the insignificant features and enhances the weight of the salient features. Extensive experiments show that our method improves the representational ability of the model and increases the robustness of the model. The contributions of our work are summarized as follows:

- Different from designing complex network structure and combining multiple convolutional neural network models, we improve the VGG structure, introduce an attention mechanism into the framework, and propose a novel single network model called Improved VGG model with Salient Attention block (IVSA).
- We firstly adapt the original model architecture and then integrate Salient Attention Block (SAB) into the adapted model to suppress non salient features and enhance more salient features, which make the framework learn more robust emotional features in order to achieve more accurate facial expression recognition.
- Extensive experiments indicate the effectiveness and superiority of the proposed IVSA. Particularly, IVSA achieves the highest accuracy 72.44% of single network without using any extra training data.

II. RELATED WORKS

A. Convolution Neural Network

The general pipeline of FER is to extract features manually or automatically after image preprocess, and then use the features to complete the subsequent expression classification. CNN is the most commonly used deep learning technique to extract image features, which could accurately learn the image feature information [10, 11, 15]. Lecun et al. [16] propose the first convolution neural network, LeNet, which can extract the structure information in the image. Krizhevsky et al. [15] propose AlexNet. This method deepens the network structure, which can learn the deeper and better dimensional feature information in the image.

Meanwhile, it also introduces the dropout mechanism to prevent the model from over fitting. Simonyan et al. [10] use the stacking method of convolution kernels, which makes the network structure deeper. ResNet [11] solves the degradation of the deep neural network through residual learning. He et al.[17] advance the regularization module in residual learning to further optimize the residual neural network.

B. *Facial Expression Recognition*

Facial expression recognition refers to the recognition of basic emotions based on facial feature information, which plays an important role in human-computer interaction [5, 6]. With the development of computer vision, facial expres-sion recognition has achieved remarkable emotion recognition accuracy. Liu et al. [18] train the three different CNNs and ensemble them to to complete FER. Minaee et al. [19] use convolutional neural network with global spatial attention module to achieve high accuracy and performance. Tang et al. [20] utilize a support vector machine to replace the softmax layer in the neural network. Pramerdorfer et al. [13] compare the performance of three different architectures, such as VGG, Inception, and ResNet. Khaireddin et al. [21] adopt the structure of VGG, and improve the accuracy of the model without using any additional training data through the adjustment of hyper parameters and the selection of optimization methods.

III. METHOD

A. *Overview*

We propose a deep learning-based FER method called Improved VGG model with Salient Attention block (IVSA). As shown in the Fig. 1, our proposed model, IVSA, contains a feature extractor and a classifier. The final output of the model is used for FER and classification.

B. *Backbone Architecture*

VGG model is a classical convolutional neural network, which has achieved good performance in multiple transfer learning tasks. It is the mainstream method to extract features from the image. The VGG network utilizes the stack of small convolution kernels to replace the large convolution kernels in the original convolution neural network, such as using two 3×3 kernels to replace one 5×5 kernel. The model not only guarantees the same scale receptive field, but also has a deeper network structure and more channels, which enables the model to extract more abundant features. Besides, VGG uses a smaller size of pooling kernel, which also enables the framework to capture more detailed feature information The reasons described above are why we choose the VGG model as the baseline.

Considering the resolution of the FER2013 dataset [22], we refer to the structure in the [21] and adjust the original VGG model. Each stage in feature extractor is composed of two convolutional modules and a max-pooling layer. The convolutional module contains a convolution layer, a activation function, and a batch normalization layer. In detail, all the kernel size in convolution layer is 3×3 and the Gaussian Error Linear Unit (GeLU) [23] is used as the activation function. Besides, both batch normalization [24] and dropout [15] are used to solve the overfitting problem,

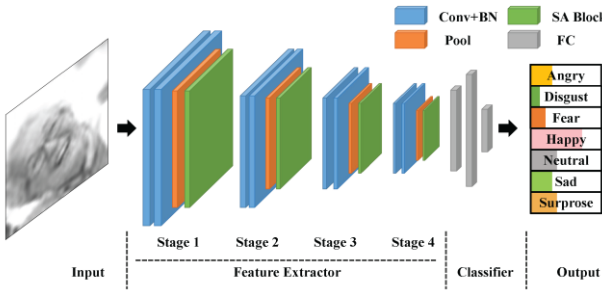

Fig. 1. **The architecture of the Improved VGG model with Attention (IVSA).** The proposed model uses images as input, and finally outputs the probability of expression labels.

speed up the learning process, and avoid gradient vanishing or explosion. The detailed operation are described as follow. Given the input feature $\mathbf{F}_{i-1} \in \mathbb{R}^{C\times H\times W}$, the stage i will generate the output $\mathbf{F}_i \in \mathbb{R}^{2C\times H\times W}$. Mathematical expression of the stage i is given as:

$$\mathbf{F}_{m_i} = \mathrm{GeLU}(\mathrm{BN}(\mathrm{Conv}(\mathbf{F}_{i-1}))) \tag{1}$$

$$\mathbf{F}_i = MaxPool(\mathbf{F}_{m_i}) \tag{2}$$

where Conv denotes a convolution layer with a 3×3 conv kernel, BN denotes batch normalization layer, GeLU denotes the Gaussian error linear unit, and MaxPool denotes a pooling layer with a 2 × 2 pooling kernel.

C. *Salient Attention Block*

We present a Salient Attention Block (SAB) to achieve non salient features suppression and more salient features enhancement, as shown in Fig.2. Inspired by visual attention [25] , we fine-tune the original channel and spatial attention submodules, respectively. Then, the SAB is inserted at the end of each stage in the feature extractor, as shown in Fig. 1. Given an intermediate feature map $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$ as input, SAB sequentially infers a channel-wise weight factor $W_c \in \mathbb{R}^{C\times 1\times 1}$ and a spatial-wise weight factor $W_s \in \mathbb{R}^{1\times H\times W}$. The overall attention process can be described as:

$$\mathbf{F}' = W_c(\mathbf{X}) \otimes \mathbf{X}, \tag{3}$$

$$\mathbf{O} = W_s(\mathbf{F}') \otimes \mathbf{F}', \tag{4}$$

where $\otimes$ denotes element-wise multiplication, and O is the final refined output. The process details of each note module are described below.

1) *Channel Attention Block*

Different from the original CAM module, we select three different types of channel aggregation features F_c^{avg}, F_c^{max} and $F_c^{L_2}$ as inputs to the channel attention module, which denote average-pooling feature, max-pooling feature and L_2-pooling feature. All features are used to generate the initial channel weights. The formula is described as follows: $\mathbf{F}_c^{avg} = AvgPool(\mathbf{F}_c)$, $\mathbf{F}_c^{max} = MaxPool(\mathbf{F}_c)$, and $\mathbf{F}_c^{L_2} =$

$L_2Pool(\mathbf{F}_c)$. Then, we get W_c by $W_c = \sigma(MLP(\mathbf{F}_c^{avg} + \mathbf{F}_c^{max} + \mathbf{F}_c^{L_2}))$

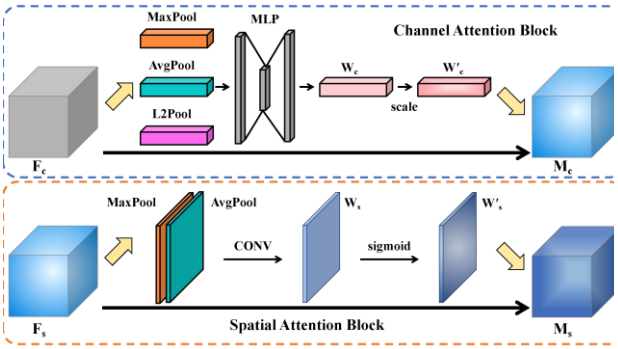

Fig. 2. **Detailed structure diagram of attention module in Salient Attention block.** As illustrated, the Channel Attention Block utilizes max-pooling outputs, average-pooling outputs and L2-pooling outputs with a shared MLP. The Spatial Attention Block utilizes similar max-pooling outputs and average-pooling along the channel axis and forward them to a convolution layer. In addition, we introduce the coefficient γ to scale the weight factors.

As proved in [26], The scaling factor γ in BN could measures the variance of channels and indicates their importance. Therefore, we introduce γ into SAB module to re-scale the weight coefficient again to improve the score corresponding to the salient features. The final output weight M_c are expressed as follows:

$$W_c' = \frac{\gamma_i}{\sum_j \gamma_j} \cdot W_c \tag{5}$$

$$M_c = W_c' \cdot \mathbf{F}_c \tag{6}$$

2) Spatial Attention Block

Referring to [25], we first apply average-pooling and max-pooling operations along the channel axis and then we use the concat operation to compose the spatial feature descriptor. After that, we apply a convolution layer with a 3×3 kernel to generate a spatial attention map W_s which indicate the salient features. Finally, we get the output features M_s and the formula is described as follows:

$$\mathrm{F_s^{avg}} = \mathrm{AvgPool}(\mathrm{F_s}), \mathrm{F_s^{max}} = \mathrm{MaxPool}(\mathrm{F_s}) \tag{7}$$

$$\mathrm{F_s} = \sigma(\mathrm{f^{7\times7}}(\mathrm{F_s^{avg}}+\mathrm{F_s^{max}})) \tag{8}$$

$$\mathrm{M_s} = \mathrm{W_s'} \cdot \mathrm{F_s} \tag{9}$$

where σ denotes the sigmoid function and $\mathrm{f^{7\times7}}$ represents a convolution operation with the kernel size of 7×7.

D. Classifier

In the classifier, the fully connected layers account for most of the parameters of the model, so we just use the three fully connected layers as the classifier and reduce the number of channels of the original model from 4096 to 1000, greatly reducing the number of parameters. Finally, we use the softmax function to get the probability of each emotion class, which is defined as follows:

$$\mathbf{O} = \mathrm{softmax}(\mathrm{FC_3}(\mathrm{FC_2}(\mathrm{FC_1}(\mathbf{X})))) \tag{10}$$

where FC_i denotes the i-th fully connected layer, **X** denotes the input features, and **O** denotes the the probability of each class.

IV. Experimental Setup

A. Dataset

On the issue of datasets selected, We evaluate IVSA on two facial expression benchmarks: FER2013 [22] and CK+ [27]. For FER2013, we use the official training, validation, and testing sets annotated in [22]. For CK+, we divide the dataset to train, valid and test according to 6:2:2. FER2013 is the most common dataset in FER task. It consists of 35,685 examples of 48x48 pixel gray scale images of faces. All the images are annotated with seven emotion categories, specifically *Angry*, *Disgust*, *Fear*, *Happy*, *Sad*, *Surprise* and *Neutral*. CK+ is an extension of Cohn-Kanade dataset, which also is commonly used in facial expression recognition research. It collects the frontal facial expressions of 123 people, and contains a total of 593 images.

B. Data Augmentation and Hyper parameters

To learn more salient features and improve the representation ability and robustness of the model, we apply several data augmentation methods in training period. Data augmentation methods include re-scale, shift horizontally or vertically, rotate, and ten-cropped. The probability that each operation is performed is 0.5. Finally, we normalize each crop images.

In the experiment, all models are trained on two Nvidia Tesla V100 GPUs. The development environment for our model is Pytorch framework based on Python. Meanwhile, we select Stochastic Gradient Descent (SGD) [28] with Nesterov Momentum and Reduce Learning Rate on Plateau (RLRP) learning rate scheduler to train the model. Besides, we set the initial learning rate as 0.01, momentum as 0.9, weight decay as 0.0001, and epochs num as 100. Finally, we choose the cross-entropy loss to optimize, and use accuracy and confusion matrix to evaluate the model.

V. Experiments

A. Comparative experiment

The experimental results of the accuracy on FER2013 are shown in Table I. We compare our method with other similar convolutional neural network models. For the baseline model VGG, the accuracy are 64.60%. GoogleNet, Inception report accuracy of 65.2 % and 71.06%, respectively. Note that the accuracy of ResNet is only 62.8. This may be that the low resolution of the input image results in poor performance. For our method, the improved model has achieved an accuracy of 70.3%. When SAB is introduced, IVSA achieves the highest accuracy of 72.44% In conclusion, it can be seen that the attention module is introduced to extract effective information on channel-wise features and spatial features, suppress the useless features, and make the model focus more on the key parts of facial expression. Furthermore, we evaluate our model on CK+ dataset to further verify the generality, generalization and robustness of IVSA, as shown in Table III.

TABLE I. ACCURACY OF DIFFERENT METHODS ON FER2013 DATASET.

METHODS	Accuracy(%)
VGG [13]	64.6
GoogleNet [12]	65.2
Resnet18 [29]	62.8
Inception [13]	71.06
IVSA w/o SAB	70.30
IVSA(Ours)	72.44

TABLE II. ACCURACY OF DIFFERENT METHODS ON CK+ DATASET.

METHODS	Accuracy(%)
LDL-ALSG [30]	93.08
IPA2LT [31]	91.67
AGRA [32]	85.27
IVSA w/o SAB	93.03
IVSA(Ours)	95.52

B. Qualitative Analysis

Visualizing the features captured inside model is important in evaluating the model and could help us to understand how model computes each class probability. For the qualitative analysis, we apply the salient map [20] to IVSA using images from the testing set. Salient map uses gradients in order to calculate the importance of the spatial locations and show the attended regions clearly. More highlight the pixels are, the most impact on the loss value they has, which helps describe how the model differentiates and captures salient features.

We generate the salient map using our IVSA model to understand how it classifies each emotion in the FER2013 dataset. Figure 3 shows salient maps for each emotion class. Obviously, our IVSA can effectively capture more salient features. The model is placing more attention on almost all facial features of the person in each image. This is most clearly seen in second column where the salient map almost perfectly maps the eyes and tooth of the man. Our model also surpress effectively less salient features like the black hair and the background, which are not salient features for emotion recognition. These saliency maps prove that our method can extract the salient features of different categories of emotions.

The ablation experiment results on FER2013 are shown in Table III. The accuracy of imporoved VGG model without SAB block is 70.3%. When we embed the attention block, such as CBAM [25], the model achieve the higher accuracy 71.72%, Improved accuracy by 1.14% compared with baseline. The improvement of accuracy indicates that the attention module can help the model select features, so as to improve the performance and performance of the model. When we adjust the attention module and replaced it with SAB, the accuracy is improved again, reaching 72.44%. Compared with other attention module, our proposed SAB module could help the model to suppress non salient features and enhance more salient features, which make the framework learn more robust emotional features in order to achieve more accurate facial expression recognition.

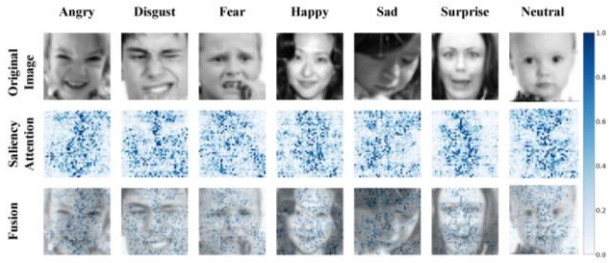

Fig. 3. **Saliency Attention Visualization for all emotion classes in FER2013.** We select a sample from each type of all emotion classes, and the ground-truth label is shown on the top of each input image.

TABLE III. ABLATION STUDY RESULT ON FER2013.

METHODS	Accuracy(%)
Improved VGG	70.30
Improved VGG + CBAM	71.72
Improved VGG + SAB(ALL)	72.44

VI. CONCLUSION

In conclusion, IVSA achieves best single-network classification accuracy on FER2013 and CK+. This paper presents an improved model with Salient Attention block for facial expression recognition. A improved framework is used as the backbone and salient attention blocks are embedded to improve the IVSA's representation ability. In addition, we introduce a re-scale factor in attention module and adjust the embed size to reduce the model parameters. IVSA in this paper achieves 72.44% and 95.52% on FER2013 and CK+ datasets, respectively. Experiment results prove the effectiveness and superiority of the proposed model. For the further works, we will explore how to further reduce the parameters of the model and how to integrate the attention module with other models.

ACKNOWLEDGMENT

This work is supported by National Key R&D Program of China (2021ZD0113502, 2021ZD0113503), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0103) and National Natural Science Foundation of China under Grant (82090052).

REFERENCES

[1] L. Shan and W. Deng, "Deep facial expression recognition: A survey," IEEE Transactions on Affective Computing, vol. PP, no. 99, 2018.

[2] X. Zhao, X. Shi, and S. Zhang, "Facial expression recognition via deep learning," Iete Technical Review, vol. 32, no. 5, pp. 347-355, 2015.

[3] M. Sajjad, M. Nasir, F. M. Allah, K. Muhammad, and S. W. Baik, "Raspberry pi assisted facial expression recognition framework for smart security in law-enforcement services," Information Sciences, vol. 479, 2018.

[4] B. Jin, Y. Qu, L. Zhang, and Z. Gao, "Research on diagnosing parkin-son ' s disease through facial expression recognition

(preprint),” Journal of Medical Internet Research, vol. 22, no. 7, 2020.

[5] G. Oh, J. Ryu, E. Jeong, J. H. Yang, and S. Lim, “Drer: Deep learning‐based driver’s real emotion recognizer,” Sensors, vol. 21, no. 6,p.2166, 2021.

[6] Y. Cui, S. Wang, and R. Zhao, “Machine learning-based student emotion recognition for business english class,” International Journal of Emerg-ing Technologies in Learning (iJET), no. 12, 2021.

[7] J. Sung, S. Lee, and D. Kim, “A real-time facial expression recognition using the staam,” in International Conference on Pattern Recognition, 2006.

[8] G. Zhao and M. Pietikainen, “Dynamic texture recognition using local binary patterns with an application to facial expressions,” IEEE Trans-actions on Pattern Analysis Machine Intelligence, vol. 29, pp. 915‐928, 2007.

[9] R. Zhi, M. Flierl, Q. Ruan, and W. B. Kleijn, “Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition,” IEEE Trans Syst Man Cybern B Cybern, vol. 41, no. 1, pp. 38‐52, 2011.

[10] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” Computer Science, 2014.

[11] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[12] M. I. Georgescu, R. T. Ionescu, and M. Popescu, “Local learning with deep and handcrafted features for facial expression recognition,” IEEE Access, 2019.

[13] C. Pramerdorfer and M. Kampel, “Facial expression recognition using convolutional neural networks: State of the art,” 2016.

[14] P. Giannopoulos, I. Perikos, and I. Hatzilygeroudis, Deep Learning Approaches for Facial Emotion Recognition: A Case Study on FER-2013. Advances in Hybridization of Intelligent Methods, 2018.

[15] A. Krizhevsky, I. Sutskever, and G. Hinton, “Imagenet classification with deep convolutional neural networks,” Advances in neural information processing systems, vol. 25, no. 2, 2012.

[16] Y. Lecun and L. Bottou, “Gradient-based learning applied to document recognition,” Proceedings of the IEEE, vol. 86, no. 11, pp. 2278‐2324, 1998.

[17] K. He, X. Zhang, S. Ren, and J. Sun, “Identity mappings in deep residual networks,” Springer, Cham, 2016.

[18] L. Kuang, M. Zhang, and Z. Pan, “Facial expression recognition with cnn ensemble,” in International Conference on Cyberworlds, 2016.

[19] S. Minaee and A. Abdolrashidi, “Deep-emotion: Facial expression recognition using attentional convolutional network,” 2019.

[20] Y. Tang, “Deep learning using linear support vector machines,” Com-puter ence, 2013.

[21] Y. Khaireddin and Z. Chen, “Facial emotion recognition: State of the art performance on fer2013,” 2021.

[22] “Challenges in representation learning: A report on three machine learn-ing contests,” Neural Networks: The Official Journal of the International Neural Network Society, 2015.

[23] D. Hendrycks and K. Gimpel, “Bridging nonlinearities and stochastic regularizers with gaussian error linear units,” 2016.

[24] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in JMLR.org, 2015.

[25] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, “Cbam: Convolutional block attention module,” Springer, Cham, 2018.

[26] Y. Liu, Z. Shao, Y. Teng, and N. Hoffmann, “Nam: Normalization-based attention module,” 2021.

[27] 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 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, 2010, pp. 94 – 101.

[28] K. L. Chung, “On a stochastic approximation method,” The Annals of Mathematical Statistics, vol. 25, no. 3, pp. 463 – 483, 1954.

[29] P. Guo and C. Song, “Facial expression recognition with squeeze-and-excitation network,” in 2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP), 2022, pp. 962 – 967.

[30] S. Chen, J. Wang, Y. Chen, Z. Shi, X. Geng, and Y. Rui, “Label distribution learning on auxiliary label space graphs for facial expression recognition,” in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13 981 – 13 990.

[31] J. Zeng, S. Shan, and X. Chen, “Facial expression recognition with in-consistently annotated datasets,” in Computer Vision – ECCV 2018,V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 227 – 243.

[32] Y. Xie, T. Chen, T. Pu, H. Wu, and L. Lin, Adversarial Graph Representation Adaptation for Cross-Domain Facial Expression Recognition. New York, NY, USA: Association for Computing Machinery, 2020, p. 1255 – 1264. [Online]. Available: https://doi.org/10.1145/3394171.3413822