2021 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-4115-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICIP42928.2021.9506181

# LEARNING ASSOCIATIVE REPRESENTATION FOR FACIAL EXPRESSION RECOGNITION

*Yangtao Du, Dingkang Yang, Peng Zhai, Mingchen Li, Lihua Zhang**

The Institute of AI and Robotics, Fudan University, Shanghai, China
JiHua Laboratory, Foshan, China

## ABSTRACT

The main inherent challenges with the Facial Expression Recognition (FER) are high intra-class variations and high inter-class similarities, while existing methods pay little attention to the association within inter- and intra-class expressions. This paper introduces a novel Expression Associative Network (EAN) to learn association of facial expression, specifically, from two aspects: 1) associative topological relation over mini-batch is constructed by similarity matrix with an adjacent regularization, and 2) learning association of expressions with Graph Convolutional Network (GCN). Besides, an auxiliary module as invariant feature generator based on Generative Adversarial Networks (GAN) is designed to suppress pose variations, illumination changes, and occlusions. Results on public benchmarks achieve comparable or better performance compared with current **state-of-the-art** methods, with 90.07% on FERPlus, 86.36% on RAF-DB, and improve by **3.92%** over **SOTA** on synthetic wrong labeling datasets.

***Index Terms***— Facial expression, Associative learning, adjacent regularization, invariant feature generator, robust representation.

## 1. INTRODUCTION

Facial expression is one of the most natural and universal means of conveying human emotional information. In computer vision community, Facial Expression Recognition (FER) help computers understand human behavior and interact with humans, with wide applications in human-computer interaction (HCI) and automatic driving. High intra-class variations and high inter-class similarities are inherent problems of FER, accompanied by external interferences such as: pose variations, illumination changes, and occlusions.

Extracting feature closely related to facial expressions from images over mini-batch is the first step to recognize expressions, and Convolutional Neural Networks (CNN) can draw features far superior to hand-crafted with rapid development of deep learning recently. However, existing methods mostly regard the object to be classified as parameters vector that should be trained independently, which calculate average loss over mini-batch by putting feature of each image into classifier ordinally. Pavlov [1] proved the important role of associative learning in human adaptive process. On the other hand, humans tend to make associative comparisons when discern subtle expressions of uncertainty. Due to the strong correlation between facial expressions, we should make comparative association judgment on FER, as illustrated in Figure 1. Meanwhile, correlation learning tends to be robust for noisy training sets, with false labels, because a single labeling error affect training process through topology structure instead of directly. As illustrated in Figure 2, we propose novel Expression Associative Network (EAN) to learn association of facial expression. Given a batch of images, a backbone CNN, whose parameters are updated by an invariant feature generator auxiliarily, is first used to extract expressional features. Then associative topological relation over mini-batch is constructed in the form of adjacent matrix by similarity matrix with an adjacent regularization. Further, GCN is used to associative learning.

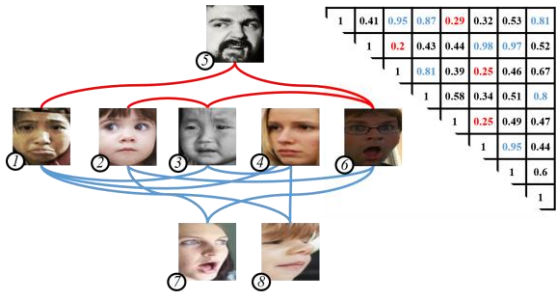


**Fig. 1.** Associative relation between images in RAF-DB dataset. The blue lines on behalf of similarity relation, while the red on behalf of opposite relation.

The algorithm performance degrades dramatically over real-world datasets due to variations introduced by pose variations, illumination changes, and occlusions. Most existing methods solve such external noise in original images' point of view. As illustrated in Figure 2, we try to give the features extracted by a backbone CNN able to resist external noise, in classified features' point of view, using GAN as auxiliary training module.

**Related work.** Generally, FER systems include three stages, face detection, feature extraction, and expression classification. Dlib[3] is used as light detector to locate faces, compared with MTCNN[4] which is more commonly used in complex scenes with better performance in profiles. For feature extraction, the stage can be divides into hand-crafted and learning-based features. Texture-based and geometry-based features consist of hand-crafted features, mainly include SIFT[5], Gabor wavelet coefficients[6], and features based on landmark points around noses, eyes, and mouths. Learning-based features always get more robust performance, as Liu *et al.*[7] pose Facial Action Units based CNN, Wu *et al.*[25] use Capsule network, and Tang *et al.*[8] propose deep CNNs winning the FER2013. Recently, Wang *et al.*[9] design regional attention networks for pose variations and occlusions, moreover, further extended to relabel images over mini-batch[10] and achieving the latest SOTA.

Works related to facial expression association are few, as the objective function of facial recognition tasks, such as FER, usually take each sample independently. Zhao *et al.*[11] utilize peak-piloted deep network (PPDN) to recognize subtle expression, with a peak expression supervising an non-peak expression during forward and backward. However, PPDN ignores the correlation between different expressions, but only considers the intensity of the same expression. Self-Cure Network (SCN) [10] is closest to our work, proposed to suppress the sample uncertainty in datasets and prevent overfitting wrong labels. However, SCN only uses self-attention weighting and relabeling images over mini-batch, without associating images to learn. Our EAN utilizes GCN to learn association between images, and it is worth to be pointed that our EAN achieve better performance on wrong labeling datasets than SCN.

Pose variations, illumination changes, and occlusions often occur in real world, Cotter *et al.*[12] use sparse classifier for FER, and Li *et al.*[13] design patch-based network to solve occlusion FER. Rudovic *et al.*[14] propose Coupled Scaled Gaussian Process Regression (CSGPR) model for head-pose normalization. Lai et al. [2] generate frontal facial images from profiles using GAN. Our EAN solves problems from the perspective of features directly, making features extracted from backbone capable to resist pose, occlusion, and illumination noise, with GAN as auxiliary module.

## 2. METHODOLOGY

Considering different expression images over mini-batch have associative relation, this paper poses Expression Associative learning Network (EAN) to explore association between images, imitating associative learning in human adaptive process.

### 2.1. Overview of EAN.

Deep features $[\mathbf{F}_0, \mathbf{F}_1, \mathbf{F}_2, \cdots, \mathbf{F}_k]$ extracted from backbone regularized by Invariant Features Generator over mini-batch.

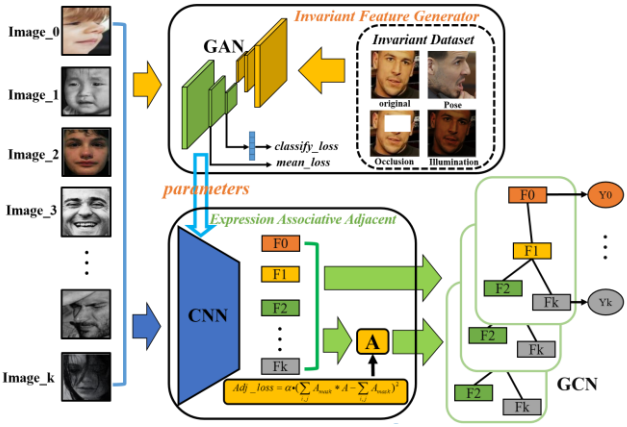


**Fig. 2.** The pipeline of our EAN. Face images are first fed into a backbone CNN for feature extraction, while Invariant Feature Generator adjust parameters in backbone. Adjacent matrix **A** is calculated and feed into GCN, with features, for associative learning.

Then we construct adjacent matrix **A** by features similarity matrix $\mathbf{C}_s$ combined with *Adj_loss*, capturing associative relation between images. Further, GCN is used to classify expression graphs based on **A** and $[\mathbf{F}_0, \mathbf{F}_1, \mathbf{F}_2, \cdots, \mathbf{F}_k]$, as shown in Figure 2.

### 2.2. Invariant Features Generator.

In order to alleviate the noise caused by pose variations, illumination and occlusion, we proposed a framework aiming to extract anti-interference feature, which incorporates a generator $G_{enc}$ and a discriminator $D_m$, as illustrated in Figure 3. The purpose of this module is to conduct confrontation learning by minimizing difference between expression features of frontal and profile images, and to promote $G_{enc}$ to generate expression features that resist the influence of pose variations, illumination and occlusion.

**Confrontation learning.** This part of inputting selects multiple frontal and profile facial images under the same identity [15]. The frontal images are respectively used to add occlusion and change the illumination intensity. Image $x$ with label $y \in [y^d, y^p]$, where $y^d$ and $y^p$ represents the label for identity and pose respectively. The feature map obtained after normalization of *Con_v3* layer in $G_{enc}$ network is selected as the input and feed to $D_m$. Difference between features of frontal and profile face judged by $D_m$, and train $G_{enc}$ under the constraint of feature classification loss to minimize the difference of facial features with the same identity.

Ensuring the features generated being related to expression, we input the frontal facial images with the expression labels $y^e$ to feed the complete $G_{enc}$. Driven by the expression classification loss, while also smoothing the

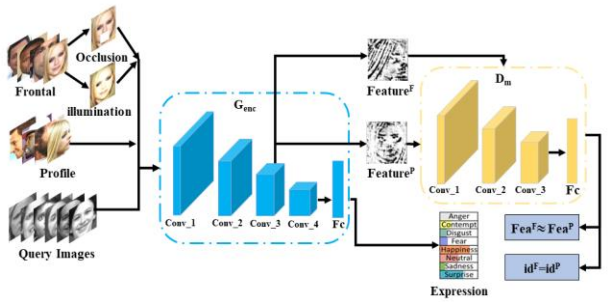


**Fig. 3.** Invariant Features Generator based on co-training of confrontation learning and expression recognition

influence of external interference, $G_{enc}$ is more focused on the extraction of expression features.

**Optimization tactics.** When the frontal and profile facial image features with different degrees of occlusion and illumination factors are feed to $D_m$, $G_{enc}$ is trained by minimizing the loss between facial features in different poses, so the optimized value function of $G_{enc}$ is expressed as:

$$\min_G V_G(D,G) = E_{x,y\sim p_{data}(x,y)} \log D^f_{y^p}(x) \quad (1)$$

Where, $D_m$ is a mutli-task CNN consisting of two parts: $D$=[$\mathrm{D}^d$,$\mathrm{D}^f$] and $D^d \in R^{M^d+1}$, $D^f \in R^{M^f}$, with $M^{d+1}$ is for identity classification, and $M^f \in [fea^F, fea^P]$ is used to characterize the features of the acquired frontal and profile images. $D_m$ is designed to estimate the identity of the sample $x$ while judging different image features obtained by $G_{enc}$ to express inconsistencies. Therefore, the optimized value function of $D_m$ is written as,

$$\max_D V_D(D,G) = E_{x,y\sim p_{data}}[\log D^d_{y^d}(x) + \log(1 - D^f_{y^p}(x))] \quad (2)$$

### 2.3. Expression Associative learning based on GCN.

*2.3.1. Expression Adjacent Matrix*

Graph Convolutional Networks (GCN) learn the topological features of information transfer between nodes through adjacent matrix, which usually be pre-defined in graph learning problems. Thus, how to build the adjacent matrix **A** is a crucial problem when adjacent relation is not provided in FER problems. According to human associative learning mechanism [1], we will correlate object with similar or opposite abstract feature. It also be proved by Thomas *et al.*[16] arguing that connected nodes in the graph are likely to share the same label. Thus, we construct associative relation in the form of an adjacent matrix over mini-batch by data-driven methods.

Images over mini-batch will transformed to feature vectors $[\mathbf{F}_0, \mathbf{F}_1, \mathbf{F}_2, \cdots, \mathbf{F}_k]$ through backbone described above. We calculate cosine similarity matrix $\mathbf{C}_s$,

$$\mathbf{C}_s(i,j) = \frac{\mathbf{F}_i \bullet \mathbf{F}_j}{\| \mathbf{F}_i \| \| \mathbf{F}_j \|}, (i,j = 0,1,\ldots,k) \quad (3)$$

Specifically, we use threshold $\tau_1, \tau_2 (\tau_2 < \tau_1)$ to filter noisy edges, and the operation can be written as,

$$\mathbf{A}(i,j) = \begin{cases} 1, \text{ if } \mathbf{C}_s(i,j) \geq \tau_1 \\ 0, \text{ if } \tau_2 < \mathbf{C}_s < \tau_1 \\ 1, \text{ if } \mathbf{C}_s(i,j) \leq \tau_2 \end{cases} \quad (4)$$

Which can also be simplified as $\mathbf{A}(i,j) = \begin{cases} 1, \text{ if } \mathbf{C}_s(i,j) \geq \tau \\ 0, \text{ if } \mathbf{C}_s(i,j) < \tau \end{cases}$.

For further optimize **A**, we consider features with same labels should be connected in graph, and the corresponding positions must be 1 in **A**. So, the *Adj_loss* is posed to partly regularize the matrix **A**. Specially, mask matrix $\mathbf{A}_{mask}$ can be constructed according to labels over mini-batch,

$$\mathbf{A}_{mask}(i,j) = \begin{cases} 1, \ label_i = label_j \\ 0, \ label_i \neq label_j \end{cases} \quad (5)$$

The Adj_loss can be written as,

$$Adj\_loss = \alpha \bullet (\sum_{i,j} \mathbf{A}_{mask} * \mathbf{A} - \sum_{i,j} \mathbf{A}_{mask})^2 \quad (6)$$

Where $\alpha$ is weight coefficient, $*$ is Hadamard product.

*2.3.2. Graph Convolutional Network (GCN)*

Graph convolutional networks are commonly used, whose basic idea is to update the representation by spreading information between nodes. The flow of information in topology diagram is similar to human associative learning, so we use GCN to explore expression association based on adjacent constructed above.

In order to reduce the complexity of the model, we define graph convolution in second order Chebyshev expansion according to Thomas *et al.*[18],

$$g_\theta \star \mathbf{x} \approx \boldsymbol{\theta} \left( \mathbf{I}_N + \mathbf{D}^{-\frac{1}{2}} \mathbf{A} \mathbf{D}^{-\frac{1}{2}} \right) \mathbf{x} \quad (7)$$

Where, **A** is adjacent matrix, $\mathbf{D}_{ii} = \sum_j \mathbf{A}_{ij}$, $\boldsymbol{\theta}$ is parameters of GCN, and **x** is feature.

We can simplify $\mathbf{I}_N + \mathbf{D}^{-\frac{1}{2}} \mathbf{A} \mathbf{D}^{-\frac{1}{2}}$ to $\tilde{\mathbf{D}}^{-\frac{1}{2}} \tilde{\mathbf{A}} \tilde{\mathbf{D}}^{-\frac{1}{2}}$, where $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}_N$ and $\tilde{\mathbf{D}}_{ii} = \sum_j \tilde{\mathbf{A}}_{ij}$, finally, the graph convolution can be written as,

$$\mathbf{Z} = \tilde{\mathbf{D}}^{-\frac{1}{2}} \tilde{\mathbf{A}} \tilde{\mathbf{D}}^{-\frac{1}{2}} \mathbf{X} \boldsymbol{\Theta} \quad (8)$$

Thus, we can model the complex inter-relationships of the nodes by stacking multiple GCN layers. For more details, we refer interested readers to [16].

## 3. EXPERIMENTS

### 3.1. Pre-processing.

In our EAN, face images are detected and aligned by MTCNN and resized to 224×224 pixels, with flipped or clipped randomly for data enhancement. The backbone is ResNet-18[17] and pre-trained on the MS-Celeb-1M.

### 3.2. Training implementation.

We train the EAN with Tesla V100 GPU and implement with Pytorch toolbox, setting the batch size as 64 or 128. During training, we first fix **A** as the identity matrix in the first 50 epochs to avoid Adj_loss exploding, instead of constructing **A** based on similarity matrix $\mathbf{C}_s$ from the beginning. The total loss function is crossentropy added by Adj_loss where $\alpha = 0.1$ , and SGD algorithm is used as optimizer. The learning rate is initialized as 0.1 which will further be divided 10 every 20 epochs.

We validate our EAN in RAD-DB and FERPlus respectively. RAF-DB[18] contains 30,000 facial images annotated with basic or compound expressions, and in our experiments, only images with seven basic expressions are used. FERPlus[19] is extended from FER2013, collected by the Google, consisting of 28709 training images, and 3589 test images. FERPlus contains 8 classes expression including *Contempt*.

### 3.3. Results.

We compare our EAN with the results of SOTA in recent years on RAF-DB and FERPlus respectively, and the experimental results are shown in Table 1 and Table 2.

**Table 1.** Comparison with SOTA methods on FERPlus

| Method | Year | Acc. |
|---|---|---|
| SeNet50[20] | 2018 | 88.8 |
| RAN[9] | 2019 | 88.5 |
| RAN-VGG16[9] | 2019 | 89.16 |
| SCN[10] | 2020 | 89.35 |
| Our EAN | 2021 | **90.07** |

**Table 2.** Comparison with SOTA methods on RAF-DB

| Method | Year | Acc. |
|---|---|---|
| DLP-CNN[18] | 2017 | 84.22 |
| GaCNN[13] | 2018 | 85.07 |
| SCN | 2020 | **87.03** |
| Our EAN | 2021 | 86.36 |

Our EAN achieve SOTA performance on FERPlus dataset, increasing by 0.72%, and achieve comparable results on RAF-DB dataset which 1.29% better than GaCNN[13]. Figure 1. represents the associative relation between 8 images over mini-batch, where coefficients are showed on the top-right corner.

Considering a single labeling error affect training process through topology structure instead of directly, we argue that our EAN is capable to suppress wrong-labeling samples from datasets. In order to verify it, we randomly choose 10%, and 20% of training data and randomly change their labels to others. In Table 3 and Table 4, we compare performance with baseline and SCN which achieve SOTA on wrong-labeling datasets, and demonstrate the robustness of EAN with far exceeding SOTA. With noise ratio 10%, our EAN outperforms the SCN 2.87%, and 3.92% with noise ratio 20%.

**Table 3.** Comparison with SOTA methods on noise 10%.

| Method | Year | RAF-DB | FERPlus |
|---|---|---|---|
| CurriculumNet[21] | 2018 | 68.5 | - |
| MetaCleaner[22] | 2019 | 68.45 | - |
| SCN[10] | 2020 | 82.18 | 84.28 |
| Our EAN | 2021 | **84.83** | **87.15** |

**Table 4.** Comparison with SOTA methods on noise 20%.

| Method | Year | RAF-DB | FERPlus |
|---|---|---|---|
| CurriculumNet[21] | 2018 | 61.23 | - |
| MetaCleaner[22] | 2019 | 61.35 | - |
| SCN[10] | 2020 | 80.10 | 83.17 |
| Our EAN | 2021 | **84.02** | **87.03** |

### 3.4. Ablation experiments.

As shown in Figure 4, the facial feature extraction result of $G_{enc}$ .By learning facial features under different conditions, even though the input (e) has the opposite posture, stronger illumination and large area occlusion, $G_{enc}$ can still get the maximum frontal facial image representation. It also reduces the loss of local features such as eyes and mouth.

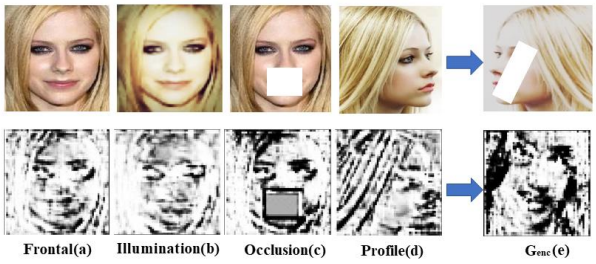


**Fig. 4.** Feature generator of $G_{enc}$ can reduce the influence of pose, occlusion and illumination

When performing expression recognition tests, the accuracy of using our trained $G_{enc}$ exceeds the baseline without it as auxiliary module, as shown in Table 5. The best performance, which improved more than 7%, fully demonstrates the effectiveness of the module.

**Table 5.** Comparison on FERPlus and RAF-DB

| | FERPlus (Acc) | | RAF-DB (Acc) | |
|---|---|---|---|---|
| Backbone | Basic | Ours | Basic | Ours |
| AlexNet[23] | 63.5 | **68.9** | 61 | **64.2** |
| VGG13[24] | 72 | **78.6** | 70.6 | **75.3** |
| VGG16 | 73.8 | **80.2** | 72.5 | **78** |
| ResNet18 | 77.6 | **82.4** | 74.1 | **81.5** |

## 4. CONCLUSION

This paper proposes a novel expression adjacent matrix and associative learning between images to explore inter- and intra-class relation in FER. The adjacent loss aims to optimize adjacent matrix to fit data. An auxiliary module works as invariant feature generator to tackle pose variations, illumination changes, and occlusions. Results on widely used datasets and synthetic wrong-labeling datasets show the effectiveness of our EAN.

## 5. REFERENCES


[1] Anderson J R, Bower G H. Human associative memory[M]. Psychology press, 2014.

[2] Lai Y H, Lai S H. Emotion-preserving representation learning via generative adversarial network for multi-view facial expression recognition[C]//2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). IEEE, 2018: 263-270.

[3] Amos B, Ludwiczuk B, Satyanarayanan M. Openface: A general-purpose face recognition library with mobile applications[J]. CMU School of Computer Science, 2016, 6(2).

[4] Zhang K, Zhang Z, Li Z, et al. Joint face detection and alignment using multitask cascaded convolutional networks[J]. IEEE Signal Processing Letters, 2016, 23(10): 1499-1503.

[5] Ng P C, Henikoff S. SIFT: Predicting amino acid changes that affect protein function[J]. Nucleic acids research, 2003, 31(13): 3812-3814.

[6] Liu C, Wechsler H. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition[J]. IEEE Transactions on Image processing, 2002, 11(4): 467-476.

[7] Liu M, Li S, Shan S, et al. Au-inspired deep networks for facial expression feature learning[J]. Neurocomputing, 2015, 159: 126-136.

[8] Tang Y. Deep learning using linear support vector machines[J]. arXiv preprint arXiv:1306.0239, 2013.

[9] Wang K, Peng X, Yang J, et al. Region attention networks for pose and occlusion robust facial expression recognition[J]. IEEE Transactions on Image Processing, 2020, 29: 4057-4069.

[10] Wang K, Peng X, Yang J, et al. Suppressing uncertainties for large-scale facial expression recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 6897-6906.

[11] Zhao X, Liang X, Liu L, et al. Peak-piloted deep network for facial expression recognition[C]//European conference on computer vision. Springer, Cham, 2016: 425-442.

[12] Cotter S F. Sparse representation for accurate classification of corrupted and occluded facial expressions[C]//2010 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2010: 838-841.

[13] Li Y, Zeng J, Shan S, et al. Occlusion aware facial expression recognition using cnn with attention mechanism[J]. IEEE Transactions on Image Processing, 2018, 28(5): 2439-2450.

[14] Rudovic O, Pantic M, Patras I. Coupled Gaussian processes for pose-invariant facial expression recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2012, 35(6): 1357-1369.

[15] Sengupta S, Chen J C, Castillo C, et al. Frontal to profile face verification in the wild[C]//2016 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2016: 1-9.

[16] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks[J]. arXiv preprint arXiv:1609.02907, 2016.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[18] Li S, Deng W, Du J P. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2852-2861.

[19] Barsoum E, Zhang C, Ferrer C C, et al. Training deep networks for facial expression recognition with crowd-sourced label distribution[C]//Proceedings of the 18th ACM International Conference on Multimodal Interaction. 2016: 279-283.

[20] Albanie S, Nagrani A, Vedaldi A, et al. Emotion recognition in speech using cross-modal transfer in the wild[C]//Proceedings of the 26th ACM international conference on Multimedia. 2018: 292-301.

[21] Guo S, Huang W, Zhang H, et al. Curriculumnet: Weakly supervised learning from large-scale web images[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 135-150.

[22] Zhang W, Wang Y, Qiao Y. Metacleaner: Learning to hallucinate clean representations for noisy-labeled visual recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 7373-7382.

[23] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[24] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[25] Wu F, Smith J S, Lu W, et al. FaceCaps for Facial Expression Recognition[C]. International Conference on Pattern Recognition (ICPR). 2020