

Facial Expression Recognition Via Multi Semantic Diffusion Model on Imbalanced Datasets

Ling Zhang¹ and Junlan Dong²

¹ Faculty of Computer, Guangdong University of Technology, Guangzhou 510006, China

² School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China
1252875930@qq.com

Abstract. This paper presents a novel facial expression recognition approach based on multiple semantic taxonomies learning on the imbalanced datasets. Recent studies on imbalanced data always concern how to homogenize the data volume between different categories, presenting strategies like minority over-sampling and majority balance cascading, etc. In this paper, we try to pay more attention in high-level semantic characterization of facial expression, using more discriminative and conceptual attributes to describe samples in the case of unbalanced sets. To fully exploit the semantic information contained in the small volume samples, we develop an Analytic Hierarchical Model (AHM) method based on facial Action Unit (AU), to enforce a discriminative mapping from the image feature space to a multi-semantic space with taxonomic relations. We apply convolutional neural networks to capture the low-level image feature, and then use dictionary learning algorithm for reconstruction of images in semantic space, in order to prevent deviation from individual identity. Experiments performed on RAF-DB, FER2013 and SFEW expression databases show that the proposed method is robust to facial expression recognition in the wild.

Keywords: semantic diffusion, imbalanced dataset, facial action control system (FACS), conceptual taxonomies, Analytic Hierarchical Model (AHM).

1 Introduction

Current studies on facial expression recognition have achieved some good results [1-4]. However, in the real world with face posture, light shade, or uneven illumination all presented multiple attributes. When smaller size of samples occurs, datasets imbalance and other factors also brought great challenges [5, 6].

The problem of imbalance in sample volume of datasets arises in semantic features learning because most datasets can be described in a series of segmented conceptual labels. An intuitive solution is Binary Relevance (BR), which trains one classifier to distinguish one semantic class from others. However, it fails to model the correlations between different conceptual labels. It is inefficient to train so many conceptual level classifiers, especially when the number of classes increases. To solve this problem, we introduce the semantic attributes of the expression image to embed low-level visual

features into an action units AUs-marked space, and in order to better learn inter-correlations among AUs, we apply an analytic hierarchical model (AHM) to enforce a discriminative mapping from the image feature space to a multi-semantic space with taxonomic relations.

2 Related Works

In aspect of imbalanced data set, Xiang et al. ^[7] utilized a weighted evaluation metric and re-sampling technique to address the imbalance issue on different pain levels datasets. Lin ^[8] introduced a practical data augmentation framework to synthesize large-scale facial images samples in the wild, combined with cluster loss to make deep features compact. Ding ^[9,31] proposed a new model, ExprGAN, which converted a face image into a series of images containing multiple expressions, and the intensity of these expressions could be continuously controlled. References [10,19] applied active learning to query a user interactively for labelling the picked examples to balance datasets volume. Dong ^[11] proposed a non-convex low-rank decomposition method combined with multiple images of the same type of expression to separate expression information from identity information. Yang ^[12] proposed a novel approach using Identity-Adaptive generator (IA-gen) to regenerate new expression images from given samples, in order to solve the data imbalance problem. In the case of keeping identity-related information unchanged, using any given facial image could create six prototypes of facial expression. Liu ^[13] employed large numbers of face images of various identities, and with their facial AUs to develop an off-the-shelf face generator for micro-expression (MiE) recognition synthesis. Cai ^[14] proposed a new Identity-free conditional Generative Adversarial Network (IF-GAN) to explicitly reduce inter-subject variation in facial expression recognition. Lan ^[30] presented a multi-region coordinate attentional residual expression recognition model (MrCAR), and by residual and multi-scale convolution networks, coordinate residual attention module was setup.

All of the above researches might not act in those datasets with only a single sample of each person, such as Real-world Affective Faces Database (RAF-DB).

Recent studies on LLM-based conversational system for the robot with social cues, were presented ^[15,19,27-29]. In addition, the semantic features are abstracted from different individuals, to solve the intra-class deviation of the expression.

Based on these approaches, we extract action units of facial images and deploy semantic diffusion model for expression recognition on imbalanced datasets. The main contributions of this paper are summarized as follows:

1. Using the AU-based hierarchical analysis model to extract the semantic attributes of facial expression images, the AU unit reflects the expression movements, which is more conducive to revealing the essence of the expression, which can alleviate the imbalance of field environment data and the impact of identity differences on recognition rate.
2. Constructing an information-shared semantic attribute vector with visual features embedded, which could enlarge inter-class difference and decrease intra-class individual variation.

3 the Proposed Method

In this section, we illustrate the proposed algorithm in details, consisting of two parts, training and testing. In training stage, X represents the visual feature of training images, and S represents the semantic attributes of training images. Embedding visual features into semantic space yields six semantic attributes, and then W can be trained from the real semantic attributes. In testing stage, the extracted visual features are embedded into the semantic space, and then the semantic feature is obtained by using the projection matrix obtained in the training phase. Finally, the angle cosine distance of the semantic features between the testing sample and each kind of expression are calculated to achieve the classification result.

3.1 Analytic Hierarchy Model based on AUs

According to the motion characteristics of different facial areas, six emotions related AUs are selected, Happy (6+12), Surprise (1+2+5+25+26), Fear (1+2+4+5+7+2), Angry (4+5+24), Disgust (9+10), Sad (1+15).

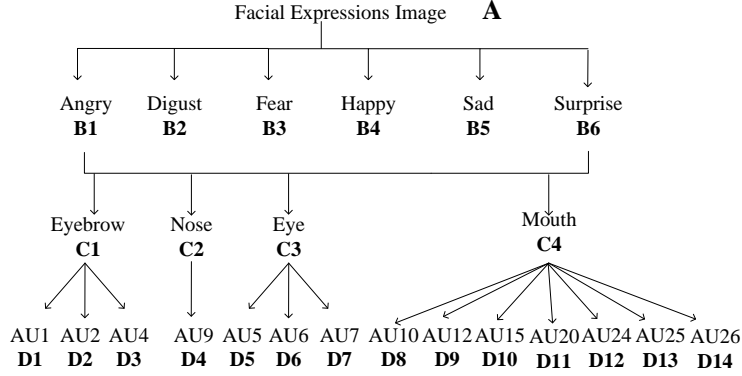


Fig. 1. Semantic hierarchy model of expressions.

Our proposed hierarchy model of facial expressions is divided into four levels, as shown in Fig.1. The first layer (A) represents the facial expression image, the second layer (B) contains six expressions (happy, disgusted, scared, angry, sad, surprise), and the third layer (C) is the four key parts of the expression face, the fourth layer (D) is the corresponding AUs. The path from the first layer to the fourth layer forms a semantic feature description, so that multiple semantic feature descriptions can represent one kind of expression. Accordingly, to describe different categories of expressions, we use a series of different weight vector combination. For example, AU6 is called "Happy Lie Detector", that is, as long as AU6 appears, the facial expression is uniquely labelled with the category of happiness. On the other side, since "surprise" and "fear", both emotions have the combination of AU1, AU2 and AU5, in order to better distinguish

between these two kinds of emotions, we should assign smaller weight value to AU1, AU2 and AU5 for the sake of homogeneity avoidance.

The semantic attributes of human face are conceptual-level features, directly reflecting the taxonomical characteristics among six kinds of facial expressions. With the semantic hierarchy model, the semantic features of a given expression image can be expressed as $S = (s_1, s_2, \dots, s_N)$, where s_i represents the i^{th} path coding weight value.

With the semantic hierarchy model, one-to-more membership between the upper and lower levels of each kind of expression is modeled. The weight values of each node in the analytic hierarchy model can be learned by reconstruction, applicable to situations where there are uncertain and class-crossover information exist. According to the importance one element upon another, we define a series of impact factor parameters, to describe the action of a layer over its previous layer, as shown in Table 1.

Table 1. Semantic Descriptions a and b

Weight value	Judgment
1	a is as important as b
3	a is slightly important over b
5	a is obviously important over b
7	a is very strongly important over b

In the following, judgment matrix R is calculated as shown in Eq. (1).

$$R = \begin{bmatrix} \frac{r_{C1}}{r_{C1}} & \frac{r_{C1}}{r_{C2}} & \frac{r_{C1}}{r_{C3}} \\ \frac{r_{C2}}{r_{C1}} & \frac{r_{C2}}{r_{C2}} & \frac{r_{C2}}{r_{C3}} \\ \frac{r_{C3}}{r_{C1}} & \frac{r_{C3}}{r_{C2}} & \frac{r_{C3}}{r_{C3}} \end{bmatrix} \quad (1)$$

Where r_i / r_j is the scale of the relative importance of the elements r_i and r_j . An example of judgment matrix of each layer of the fear image is shown in Table 2.

Table 2. Fear image's judgment matrix of each layer

- a. the discriminant matrix of B layer b. the second layer of C1 sub-discrimination matrix

B3	C1	C3	C4		C1	D1	D2	D3
C1	1	1/5	1/3		D1	1	3	1
C3	5	1	3		D2	1/3	1	1/3
C4	3	1/3	1	a.	D3	1	3	1

c. the second layer of C3 sub-discrimination matrix

C3	D5	D7
D5	1	5
D7	1/3	1

The relative weight calculation formula of each layer is as described in Eq. (2).

$$W = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} = \begin{bmatrix} \sqrt{\prod_{j=1}^n r_{1j}} \\ \sqrt{\prod_{j=1}^n r_{2j}} \\ \vdots \\ \sqrt{\prod_{j=1}^n r_{nj}} \end{bmatrix} \quad (2)$$

Then we normalize all row weights w_i to get a relative weight value, as shown in Eq. (3).

$$U_i = w_i / \sum_{i=1}^n w_i \quad (3)$$

Similarly, the relative weight values of all layers can be obtained by analogy. Finally, the semantic feature vector of an image can be obtained as described in Eq. (4).

$$S = (s_1, s_2, \dots, s_N) \quad (4)$$

N is the total number of path node codes in the hierarchical structure model, where s_i is calculated as in Eq. (5).

$$s_i = U_i^{(1)} \times U_i^{(2)} \times U_i^{(3)} \quad (5)$$

Where $U_i^{(1)}$ represents the relation weight value of layer B, and $U_i^{(2)}$ represents that of layer C, while $U_i^{(3)}$ represents that of layer D.

3.2 Embedding of the Semantic Space

When samples number of some categories are very small in facial expression recognition, the semantic attributes of restricted samples can share the information of emotional features. On the other hand, when samples differ from each other significantly within a same category, common features should be summarized and derived out these few samples. Image features extracted by deep networks (such as Resnet) make up of low-level visual feature spaces, and semantic feature representation S of the expression images consist of their semantic space. Learning a semantic self-encoder can get a projection function from visual feature space to semantic space.

Suppose X represents visual feature, and S are the attribute features. We define X' as the reconstructed visual feature vectors.

Let the projection matrix from the visual layer to the semantic layer be W , and the projection matrix from the semantic layer to the reconstructed visual layer be W^T . In order to make the input and output as the same as possible, we give a reconstruction objective function, while constraining the S layer representation errors at the same time, as followed in Eq. (6).

$$\min_W \|X - W^T S\|_F^2 + \lambda \|WX - S\|_F^2 \quad (6)$$

Considering the matrix property $\text{Tr}(X) = \text{Tr}(X^T)$ and $\text{Tr}(W^T S) = \text{Tr}(S^T W)$, the Eq. (6) is converted into Eq. (7).

$$\min_W \|X^T - S^T W\|_F^2 + \lambda \|WX - S\|_F^2 \quad (7)$$

Finally, with the derivative setting to 0, the Eq. (8) is achieved from Eq. (7).

$$-S(X^T - S^T W) + \lambda(WX - S)X^T = 0 \Rightarrow SS^T W + \lambda WXX^T = SX^T + \lambda SX^T \quad (8)$$

Assuming that $A=SS^T$, $B=\lambda XX^T$, $C=(1+\lambda)SX^T$, Eq. (8) can be expressed as shown in Eq. (9).

$$AW + WB = C \quad (9)$$

The Eq. (9) is a Sylvester equation, which can be solved using the Bartels-Stewart algorithm, and after that we get the projection matrix W .

Finally, we can embed a test sample x_{te} into the semantic space. The classification result of the test image can be returned by estimating the minimum value of the cosine angle distance between s_{te}' and the true semantic representation s_{te} in the semantic space. The calculation is as shown in Eq. (10).

$$\varphi(x_{te}) = \underset{j}{\text{argmin}} D(s', s_j) \quad (20)$$

Where s is one true semantic attribute vector of six expressions, and D is the angle cosine distance function, while $\varphi(x_{te})$ returns a classification result.

The detailed diagram of our proposed expression recognition model is illustrated as shown in Fig.2.

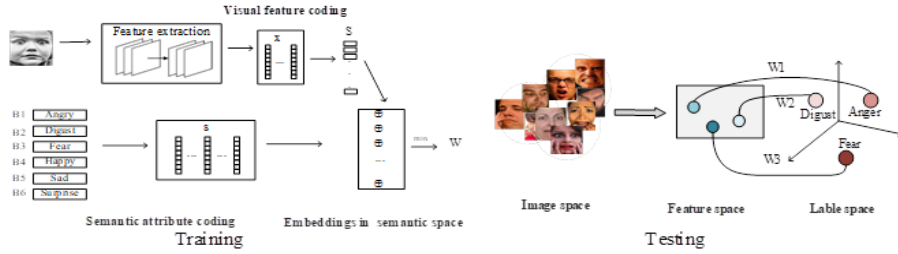


Fig. 2. The architecture of our expression recognition model.

4 Experiment Results

4.1 Experimental analysis of two schemes

In order to verify that the semantic description proposed in this paper is helpful for expression recognition, three databases FER2013, RAF-DB and SFEW are tested based on two neural networks: Resnet34 and Resnet101. Scheme1 represents Resnet34 combined with our proposed approach. Scheme2 represents Resnet101 combined with our proposed approach. The experimental results are shown in Fig.3.

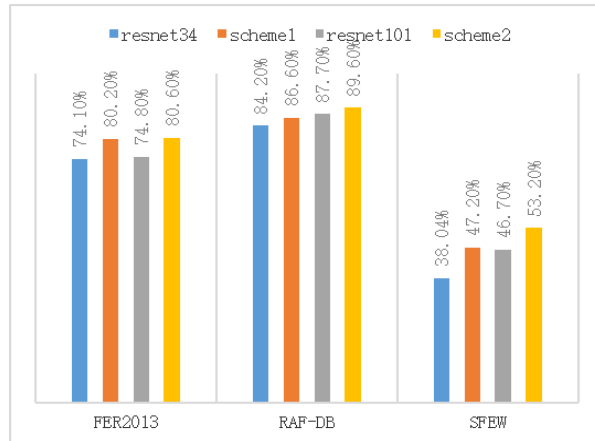


Fig. 3. Accuracy of the baseline model (Resnet and our proposed model) on three datasets.

It can be seen from Fig.3 that the deeper the network is, the more features are extracted from sample images. By the way, the more semantic information it owns, and the higher the recognition rate would be. With the help of semantic attributes, the recognition rate of FER2013 datasets with Resnet34 increased by 6.1% and with Resnet101 by 5.8%. In the RAF-DB database, Resnet34 increased its recognition rate by 2.4% and

Resnet101 by 1.9%. In the SFEW database, Resnet34 improved its recognition rate by about 9.2% and Resnet101 by 6.5%. In three databases, due to the distortion and deformation of face appearance caused by various head deflection and occlusion, their data distribution diversity is significant, thereby the degree of their improvement in recognition rate varies differently.

4.2 Comparison of experimental results on three datasets

We have selected some of the latest methods for comparison. Tian^[18,23] proposed a concept of triple loss to improve the inter-class distance of image recognition and reduce the intra-class distance, which is the same idea as our method. Yang^[26] not only extracted the overall feature information of the face but also extracted the local feature information of the eyes and mouth region based on CNN and attention mechanism and fuses the output for facial emotion recognition.. Zhang^[16] proposed self-similarity learning based on convolutional neural networks with small inputs. Liu^[17] aimed to address the lack of large-scale datasets in micro-expression (MiE) recognition, and replaced these areas with corresponding AU units in combination with FACES coding. However, they need large numbers of face images from various identities to yield MiE dataset. With small samples, we use the method of hierarchical analysis to accurately extract the AU unit of each expression, and assign different weights to different AU units of the same kind of expression to better reflect the detailed information of the expression. Experimental comparison is as shown in Table 3.

Table 3. Accuracy of our model regarding several methods on three datasets

Methods	FER2013	RAF-DB	SFEW
Yang[26]	71.8%	85.13%	--
Liu [17]	72.1%	--	--
Tian[23]	72.64%	--	--
MPCSAN[25]	--	74.20%	51.05%
Hua[21]	--	76.73%	47.43%
Liu [24]	--	73.19%	46.1%
Yang[26]	71.47%	85.2%	52.75%
Ours	80.6%	89.6%	53.2%

In Table 3, it can be seen that the recognition rate of our method is superior to others, with an obvious increase of about 8% in Fer2103 dataset.

Hua^[21] introduced a CNN with densely backward attention to leverage the aggregation of channel-wise attention at multi-level features in a backbone network for reaching high recognition performance with cost-effective resource consumption. However, in the wild environment, faces are multi-pose and diverse, so it is difficult to accurately extract these three regions. Taking it into consideration, a joint spatial and scale attention network (SSA-Net) was used to localize proper regions for simultaneous head pose estimation (HPE)^[24]. SSA-Net was deployed to discover the region most

relevant to the facial expression in a hierarchical scale by a spatial attention mechanism which only considered local information. we use semantic attributes on both the details of the expression image and the global information. In RAF-DB database, our method has improved by about 4% higher in recognition accuracy. Gong^[25] proposed an effective multi-head parallel channel-spatial attention network (MPCSAN) for face expression recognition in the wild, consisting of a feature aggregation network (FAN), a multi-head parallel attention network (MPAN), and an expression forecasting network (EFN). As it can be seen from Table 3, compared with this kind of combination of three networks, our proposed semantic attribute-based approach is evenly more conducive in classification accuracy.

In addition, from the comparison results in Table 3, we can see that our method can also achieve comparable or even better performance than other state-of-the-art methods, not only for RAF-DB, but also for other two databases. This indicates that our proposed method showing an excellent ability to learn relevant formation from very limited amount of data.

5 Conclusion

This paper presents the use of semantic descriptions in emotional classification tasks to solve the problem of data imbalance in the wild expression database and the diversity of same expressions. A semantic attribute for generating auxiliary data for a few classes using a hierarchical analysis model is proposed. On the one hand, in the process of establishing AHM model, the relationship between AU unit and various expressions is studied, which not only minimizes the difference within the categories, but also considers the identity between the categories of the expression images. On the other hand, we use the transfer learning method to extract the visual features of facial expressions by using neural network, so as to make the feature expression more detailed and specific. Finally, experiments on RAF-DB, Fer2013, and SFEW three benchmark datasets show that our semantic attribute-based auxiliary data technology can improve the distribution integrity and boundary clarity between classes.

Acknowledgments. This research was supported by the science and technology project of the department of transportation of Guangdong province, China (science and technology - 2016 -02-030).

References

- [1] S. Li, W.H. Deng. Deep Facial Expression Recognition: A Survey[J]. IEEE Transactions on Affective Computing, 2020, 13(3): 1195-1215.
- [2] D. Mehta, M.F.H. Siddiqui, A. Javaid. Facial Emotion Recognition: A Survey and Real-World User Experiences in Mixed Reality [J]. Sensors, 2018, 18(2): 416-416.
- [3] S.J. Ji, K. Wang, X.J. Peng, et al. Multiple Transfer Learning and Multi-label Balanced Training Strategies for Facial AU Detection in the Wild[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW): 1657-1661.
- [4] E. Sariyanidi, H. Gunes, A. Cavallaro. Learning bases of activity for facial expression

- recognition [J]. *IEEE Transactions on Image Processing*, 2017, 26(4): 1965-1978.
- [5] I. Rieger, J. Pahl, D. Seuss. Unique Class Group Based Multi-Label Balancing Optimizer[C]. 2020, the 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020): 619-623.
- [6] D.D. Deng, Z.K. Chen, B.E. Shi. Multitask Emotion Recognition with Incomplete Labels[C]. 2020, the 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020): 592-599.
- [7] X. Xiang, F. Wang, Y.W. Tan, et al. Imbalanced regression for intensity series of pain expression from videos by regularizing spatio-temporal face nets[J]. *Pattern Recognition Letters*. 2022,163: 152-158.
- [8] F. Lin, R.C. Hong, W.G. Zhou, et al. Facial expression recognition with data augmentation and compact feature learning[C]. 2018, the 25th IEEE International Conference on Image Processing (ICIP): 1957-1961.
- [9] I. Sundin, P. Schulam, E. Siivola, et al. Active learning for decision making from imbalanced observational data[C]. 2019, the 36th International Conference on Machine Learning (ICML), 97: 6046-6055.
- [10] T. T. D. Pham, C. S. Won. Facial Action Units for Training Convolutional Neural Networks[J]. *IEEE Access*. 2019, 7: 77816-77824.
- [11] J.L. Dong, L. Zhang, Y.H. Chen, et al. Occlusion expression recognition based on non-convex low-rank double dictionaries and occlusion error model[J]. *Signal Processing: Image Communication*. 2019, 76: 81-88.
- [12] H.Y. Yang, Z. Zhang, L.J. Yin. Identity-Adaptive Facial Expression Recognition through Expression Regeneration Using Conditional Generative Adversarial Networks[C]. 2018, the 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018): 294-301.
- [13] Y.C. Liu, Z.D. Wang, T. Gedeon, et al. How to Synthesize a Large-Scale and Trainable Micro-Expression Dataset? [C]. *ECCV 2022, European Conference on Computer Vision*, 8: 38-55.
- [14] J. Cai, Z.B. Meng, A. S. Khan, et al. Identity-Free Facial Expression Recognition using Conditional Generative Adversarial Network[C]. 2021, the 28th IEEE International Conference on Image Processing (ICIP): 1344-1348.
- [15] Y.K. Lee, Y.W. Jung, G. Kang, et al. Developing Social Robots with Empathetic Non-Verbal Cues Using Large Language Models[C]. 2023, the 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN): 1-5.
- [16] L. Zhang, W.C. Jiang, W. Xiang. Dictionary learning based on structural self-similarity and convolution neural network [J]. *Journal of Ambient Intelligence and Humanized Computing*. 2022.3, 13(3) SI: 1463-1470.
- [17] Y.C. Liu, Z.D. Wang, T. Gedeon, et al. How to Synthesize a Large-Scale and Trainable Micro-Expression Dataset?[C]. 2022 the 17th European Conference on Computer Vision (ECCV), 13668: 38-55.
- [18] Y. Tian, Z.W. Wen, W.C. Xie, et al. Outlier-Suppressed Triplet Loss with Adaptive Class-Aware Margins for Facial Expression Recognition[C], 2019, IEEE International Conference on Image Processing (ICIP): 46-50.
- [19] Y.L. Gan, J.Y. Chen, L.H.Xu Facial expression recognition boosted by soft label with a diverse ensemble[J]. *Pattern Recognition Letters*. 2019, 125(1): 105-112.
- [20] S. Li, W.H. Deng. Reliable Crowdsourcing and Deep Locality-Preserving Learning for Unconstrained Facial Expression Recognition[J]. *IEEE Transactions on Image Processing*. 2019, 28(1): 356-370.
- [21] C.H. Hua, T. Huynh-The, H. Seo, et al. Convolutional Network with Densely Backward Attention for Facial Expression Recognition[C]. 2020, the 14th International Conference on Ubiquitous Information Management and Communication (IMCOM). DOI: 10.1109/IMCOM48794.2020.9001686.
- [22] L. Liang, C. Lang, Y. Li, et al, Fine-Grained Facial Expression Recognition in the Wild[J]. *IEEE Transactions on Information Forensics and Security*. 2020, 16: 482-494.

- [23] L.L. Cui, Y. Tian. Facial Expression Recognition by Regional Attention and Multi-task Learning[J]. *Engineering Letters*. 2021, 29(3):
- [24] Y.Y. Liu, J.Y. Peng, W. Dai, et al. Joint spatial and scale attention network for multi-view facial expression recognition[J]. *Pattern Recognition*, 2023, 139: <http://dx.doi.org/10.1016/j.patcog.2023.109496>
- [25] W.J. Gong, Y.R. Qian, Y.Y. Fan, et al. MPCSAN: multi-head parallel channel-spatial attention network for facial expression recognition in the wild[J]. *Neural Computing and Applications*. 2023, 35(9): 6529-6543.
- [26] Y.Q. Yang, H. Zhou. A Multi-region Feature Extraction and Fusion Strategy Based CNN-Attention Network for Facial Expression Recognition[C]. *ROSENET 2022: the 6th EAI International Conference on Robotic Sensor Networks*: 67-79.
- [27] Q. Dong, W.H. Ren, Y. Gao, et al. Multi-Scale Attention Learning Network for Facial Expression Recognition[J]. *IEEE Signal Processing Letters*. 2023, 30: 1732-1736.
- [28] D.L. Chen, G.H. Wen, H.H. Li, et al. Multi-Relations Aware Network for In-the-Wild Facial Expression Recognition[J]. *IEEE Transactions on Circuits and Systems for Video Technology*. 2023, 33(8): 3848-3859.
- [29] G.B. Li, X. Zhu, Y.R. Zeng, et al. Semantic relationships guided representation learning for facial action unit recognition[C]. 2019, the AAAI Conference on Artificial Intelligence, 33(01): 8594-8601.
- [30] J.H. Lan, X.G. Jiang, G.J. Lin, et al. Expression Recognition Based on Multi-Regional Coordinate Attention Residuals[J]. *IEEE Access*. 2023, 11: 63863-63873.
- [31] H. Ding. Facial Expression Recognition and Editing with Limited Data[D]. *Dissertations University of Maryland, College Park*, 2020.