Natten-Deim Facial Expression Classification Network Based on Neighborhood Attention and Dual-Branch Feature Fusion

ORIGINALITY REPORT

10%	8%	8%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	pubmed.ncbi.nlm.nih.gov Internet Source	2%
2	www.mdpi.com Internet Source	2%
3	www.frontiersin.org Internet Source	1%
4	Hongwen Zhang, Qi Li, Zhenan Sun, Yunfan Liu. "Combining Data-driven and Model-driven Methods for Robust Facial Landmark Detection", IEEE Transactions on Information Forensics and Security, 2018 Publication	1%
5	arxiv.org Internet Source	1%
6	Danfeng Yan, Jiyuan Chen, Jianfei Cui, Ao Shan, Wenting Shi. "Deep Multi-Head Attention Network for Aspect-Based Sentiment Analysis", 2019 IEEE International Conference on Big Data (Big Data), 2019 Publication	1%
7	Submitted to King's College Student Paper	<1%
8	"A Survey of Transformers in Video Prediction", Academic Journal of Computing & Information Science, 2023 Publication	<1%
9	assets-eu.researchsquare.com	

Natten-Deim: Facial Expression Classification Network Based on Neighborhood Attention and Dual-Branch Feature Fusion

Weihong Luo[a,1], Ziyi Chen [b,2], Cheng Peng[a,3*]
[a]University of Electronic Science and Technology of China, Zhongshan Institute, Zhongshan, Guangdong, 528402, China
[b]South China Agricultural University, Guangzhou, Guangdong, 510642, China
[1]1054005418@qq.com, [2]clarakitty.45666@gmail.com
[*3]Corresponding author's e-mail: pengcheng@zsc.edu.cn

ABSTRACT

The main problems faced by facial expression recognition are: insufficient multi-scale feature extraction and low efficiency of feature fusion. To solve this problem, this paper combines the dual-branch backbone network and neighborhood attention to establish a new expression recognition model. The method constructs a dual branch structure of landmark prior branch (LPB) and expression backbone branch (EBB) to realize the output of multi-modal features at three levels. The model designs dual Feature Pyramid Module (DFPM) to achieve efficient fusion of dual branch features, and uses multi-scale adaptive neighborhood attention mechanism (NATTEN Block) to configure different sizes of receptive fields for feature maps of different scales to enhance local spatial features and adaptively focus on key regions of expression. Finally, the convolutional network was used to extract fine features, and the multi-head self-attention mechanism was used to optimize the overall features. On the Real-world Affective Faces Database (RAF-DB) dataset, the accuracy reaches 85.28%. This method verifies the effectiveness of the joint action of the neighborhood attention mechanism and the dual-branch feature fusion strategy in the task of facial expression recognition, providing valuable theoretical and practical support for the development of computer vision and affective computing.

Keywords: Facial expression classification, Two-branch backbone network, Neighborhood attention, Multi-scale feature fusion, Feature fusion.

1. INTRODUCTION

Facial expression detection needs to meet the two key goals of expression localization and fine-grained feature discrimination at the same time. Compared with the traditional target detection methods, facial expression has the inherent characteristics of small scale, small variation range and dense spatial distribution. In the real environment, there are many interferences such as position offset, partial occlusion, and unstable illumination. This puts forward strict requirements for face recognition models: It is not only necessary to accurately extract the geometric and topological relationships of important facial parts (such as eyebrows, eyes, mouth, etc.), but also to ensure the effective transfer and aggregation of discriminative features at multiple scales, such as feature pyramid network FPN 错误!未找到引用源。 and path aggregation network PAN[2]. Limited by the limited training samples and complex actual environment, only using the backbone network to learn it is easy to cause slow convergence speed of the model.

Current improvement strategies for expression detection tasks mostly focus on fine-tuning the detection head structure or local improvement of multi-scale pyramid networks (e.g., classic detection architectures such as YOLO[3], SSD[4], RetinaNet[5],etc.). Feature fusion modules such as FPN 错误!未找到引用源., Path Aggregation Network (PAN)[2], and detection methods such as DETR[6], DEIM[7]), Li Fei [8]et al. proposed Triple-ATFME network to realize efficient micro-expression recognition on multiple datasets by preprocessing key points of 68 faces and extracting TV-L1 optical flow features, and then fusing three-branch ShuffleNet with CFAM module. Hafiz Khizer Bin Talib[9] et al. proposed ConVAT model, which combined CNN and multi-head attention mechanism, and verified by LOSO, achieved high-accuracy micro-expression recognition on SAMM, CASME II and SMIC datasets. However, these methods generally lack a clear and stable geometric alignment transmission path, which makes it difficult for the model to accurately direct attention to the core correlation area of expression discrimination in the initial stage of feature learning, thus limiting the effective mining and representation of key features. This type of model lacks a clear and stable feature alignment path, which leads to the inability to accurately guide attention to the relevant areas of facial expression recognition in the early stage of training and learning, reducing the efficient extraction of features. In the test process of RAF-DB dataset, the traditional YOLO

series models have obvious shortcomings in the recognition performance of specific expression categories. Among them, the AP50 index of YOLOv3 model for "disgust" expression only reaches 28.56%, and the AP50 value for "fear" expression detection is only 37.02%. What is more prominent is that the YOLOv4 model directly reduces the detection accuracy to 0% in the recognition task of "disgust" expression. These measured data show that the existing expression detection methods have significant shortcomings in recognition ability when dealing with expression categories with small number of samples and easily disturbed. The AP50 of the baseline model for "angry" expression has reached 89.80%. From the perspective of actual application requirements, there is still room for further optimization and improvement of this index. In addition, the hybrid mode of "channel stacking" commonly used in the[7] (CCFF) constructed based on the DEIM algorithm is easy to add redundant mutual exclusive information in the feature processing, which reduces the identification of strong channel signals and reduces the robustness of the model in complex environments.

Therefore, current expression recognition methods face two major bottlenecks: 1) lack of reliable geometric prior constraints (such as facial keypoint constraints, face recognition, etc.); Existing algorithms include ArcFace[10], FAN[11], etc. Second, the filtering mechanism of redundant information at the channel level has not been constructed yet. These two problems are the bottlenecks that restrict the performance improvement of face recognition algorithms.

Therefore, this paper proposes a NATTEN-DEIM model, which follows the design principle of "dual-branch feature extraction + adaptive attention modulation + progressive multi-scale fusion". The model adopts a dual-branch parallel architecture of landmark prior branch and expression backbone branch, and extracts multi-modal features synchronously at three scale levels, which not only provides accurate and stable geometric prior information for expression recognition tasks, but also outputs rich semantic feature representations. A dual-gated facial prior modulator was designed to adaptively select and fuse the complementary features of the dual branches based on the gating mechanism to dynamically enhance the key regions of expression. A multi-scale adaptive neighborhood attention module is introduced to configure receptive fields of 3×3, 5×5 and 7×7 for feature maps of different scales, respectively, to focus on local areas with significant expression changes. The multi-head self-attention mechanism is combined to model global feature dependencies, which further optimizes the quality of multi-scale fusion features and improves the overall performance of the model.

2. METHOD

2.1 Overall network architecture

The Atten-DEIM model constructed in this paper is based on the accuracy of the facial expression recognition task. The effect of expression recognition significantly depends on the geometric structure correlation of key regions such as eyes, eyebrows, and mouth. Therefore, the dominant key point prior branch (LPB) was introduced to guide the network to accurately focus on such structural core regions in the initial stage of training by relying on stable facial key point prior information. Aiming at the problems of redundant information and inter-channel interference caused by traditional multi-scale feature fusion using stacking method, this paper adopts double-gate facial prior modulator (DFPM) and multi-scale neighborhood attention module (NATTEN), following the technical idea of "prior guidance to achieve controllable modulation and selective multi-scale fusion". Strengthening the local pattern representation which has substantial value for expression discrimination.

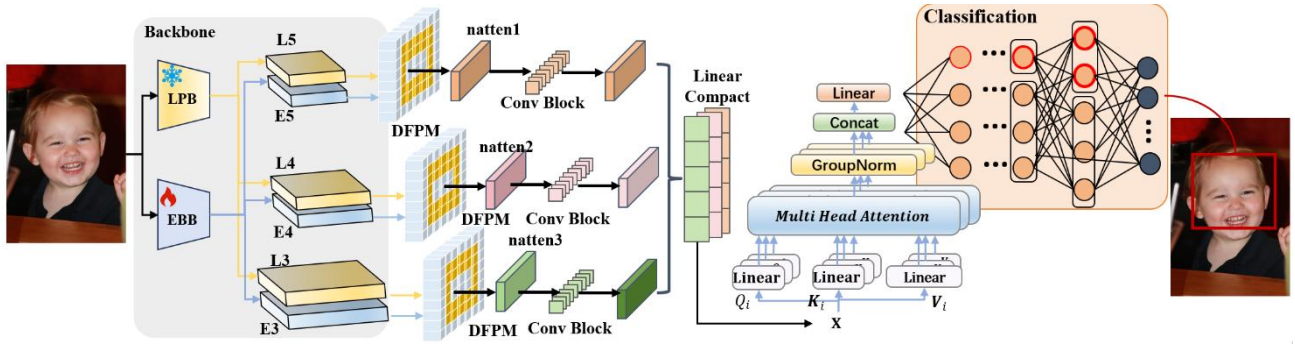

Figure 1.Overall network architecture(The sample facial images are from the Real-world Affective Faces Database (RAF-DB)[13]).

As shown in Figure 1, the model builds a dual-branch backbone network architecture, and sends the input face image into the landmark Prior branch (LPB) and the expression backbone branch (EBB) respectively to extract the features of L5/E5,

L4/E4, and L3/E3 three scale levels. The dual-gated facial prior modulator (DFPM) was used to fuse the features output by LPB and EBB at each scale level, and the multi-scale adaptive Neighborhood Attention module (NATTEN) was used to implement feature enhancement. In this process, different scale feature maps are matched with different receptive field sizes: L5/E5 level uses 3×3 convolution kernels, L4/E4 level corresponds to 5×5 convolution kernels, and L3/E3 level is configured with 7×7 convolution kernels to realize adaptive focusing of facial expression regions. The output features of each NATTEN module are further optimized by the convolutional layer, and the features of the three scales are uniformly mapped to the same dimension, and the alignment operation of the spatial size is completed, and finally spliced and fused. The fused multi-scale features are used to build dependencies through the multi-head attention mechanism, and finally output by the fully connected layer to complete the classification task of seven types of expression.

2.2 Double Gate Surface Prior modulator (DFPM)

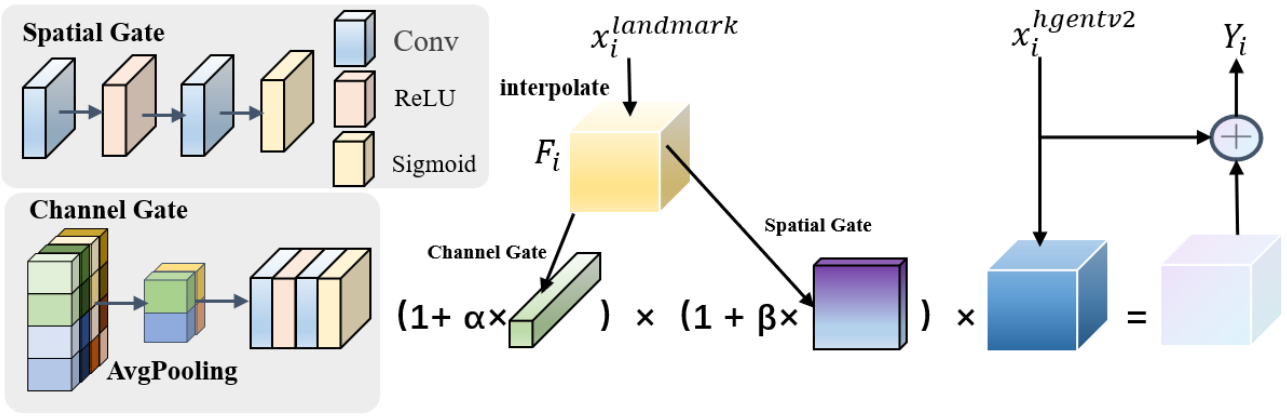

Figure 2.Double Gate Surface Prior modulator (DFPM).

The core design of DFPM is to integrate the geometric prior information extracted by the LPB module into the semantic features output by the EBB module in a multi-scale feature level in a strategy that is both controllable and interpretable (as shown in Figure 2). For the scale index $i \in \{3,4,5\}$, we define the geometric prior after alignment as F_i, and the features of the LPB branch are denoted as X_i^{landmark}. However, the semantic features of the EBB branch are denoted as X_i^{hgnetv2}. This is implemented as follows: InterpolateX_i^{hgnetv2}so that its spatial resolution exactly matches X_i^{hgnetv2} to obtain the intermediate feature map $\tilde{F}_i$. With the help of two gated subnetworks that are independent of the semantic feature stream, The channel gating and space gating are constructed from $\tilde{F}_i$ to realize the precise control of X_i^{hgnetv2}.

- Channel gating: The prior feature $\tilde{F}_i$ of this module is used as the core input basis, and the importance weight is assigned to the channel dimension of the semantic feature. Generate the weight tensor ch_{gate} whose dimensions satisfy $\mathbb{R}^{B\times C_g\times 1\times 1}$. The acquisition of weight needs to combine the processing method of global spatial aggregation and the low-rank modeling strategy of correlation characteristics between channels to accurately define the feature channels that should be strengthened or weakened, and maintain the consistency of semantic information in deeper feature levels.
- Space gate control: The prior feature $\tilde{F}_i$ is used as the core input reference, and the corresponding attention feature map is constructed for the spatial dimension. The final output dimension satisfies the attention tensor $\mathbf{sp}_{\text{gate}}$ of $\mathbb{R}^{B\times C_g\times 1\times 1}$. With the help of the local context modeling method, a smooth and continuous attention weight distribution can be constructed to locate the key area in the feature map, and the target contour and detail information at the high resolution level can be strengthened.

The above two types of gating mechanisms combine the learnable scaling parameters α and β, which are initially set to 0, and act on the semantic feature X_i^{hgnetv2}. The corresponding mathematical relationship is as follows:

$$\tilde{X}_i^{\text{hgnetv2}} = X_i^{\text{hgnetv2}} \cdot \left(1 + \alpha \cdot \mathbf{ch}_{\text{gate}}\right) \cdot \left(1 + \beta \cdot \mathbf{sp}_{\text{gate}}\right) \tag{1}$$

Setting the initial value of α, β to 0 can make DFPM present the characteristics of approximate identity mapping in the initial stage of training, and isolate the interference of external prior on the convergence of the backbone network. As the

training process continues, α, β will adaptively increase the value, and gradually integrate the channel and spatial prior information into the semantic feature flow to achieve flexible and controllable feature control effect.

In order to achieve the purpose of suppressing noise and enhancing numerical stability, DFPM introduces a 3×3 depth-wise separable convolution of a single normalization and activation operation after feature modulation, and further constructs a residual connection structure with the original features. The corresponding mathematical relationship can be expressed as follows:

$$Y_i = \mathrm{ReLU}\left(\mathrm{BN}\left(\mathrm{DWConv}_{3\times3}\left(\tilde{X}_i^{\mathrm{hgnetv2}}\right)\right)\right) + X_i^{\mathrm{hgnetv2}} \tag{2}$$

For any dimension hierarchy i, F_i and X_i^{hgnetv2} space size $(H_i W_i)$,The dimension of $\mathbf{ch}_{\mathrm{gate}}$ satisfies $\mathbb{R}^{B\times C_g\times1\times1}$, While $\mathbf{sp}_{\mathrm{gate}}$ dimensions for $\mathbb{R}^{B\times1\times H_i\times W_i}$.

DFPM adopts the core process of "alignment first, backdoor control, light correction" to achieve a stable and progressive injection of geometric domain markers into the semantic domain. Compared with the scheme that directly incorporated prior information into the semantic stream, the module effectively alleviated the problems of statistical shift and training fluctuation by controlling the strength parameter (α/β) and decoupling the gate structure. On the basis of not changing the backbone network topology, the discrimination performance between high threshold localization accuracy and difficult samples is further improved.

2.3 Multi-head attention mechanism

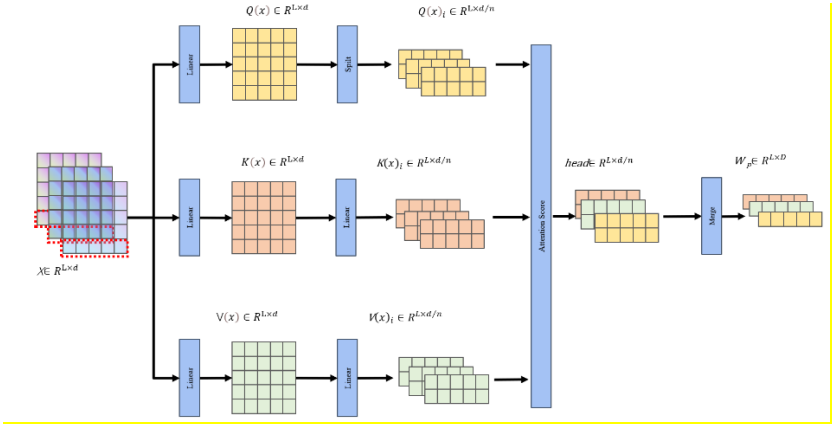

Figure 3.Head Attention Mechanism.

The input of multi-head attention mechanism consists of three core vectors, namely, query vector (q), key vector (k) and value vector (v). For any specified query vector, the mechanism will calculate the similarity values between the query vector and each vector, and construct the corresponding weight parameters. The weighted sum operation is carried out on the key vector using the obtained weights, and the operation results are fused with the value vector to output the final feature representation of the multi-head attention mechanism. The mathematical expression is as follows:

$$MultiHead(Q,K,V) = Concat(head_1, \cdots head_h)W^0 \tag{3}$$

Where Q, K, V correspond to the query vector, key vector, and value vector respectively, h represents the number of heads of the attention mechanism, $head_h$ refers to the output feature of the HTH attention head, and W^0 is the output transformation matrix used for feature mapping.By default, 8 heads are used in multi-head attention, and the input/output channel dimension is 256.

3. EXPERIMENT

3.1 Experimental Setup

The Atten-DEIM model proposed in this study is built based on the PyTorch framework. In its dual-branch backbone network architecture, the key point prior branch (LPB) is constructed by the pre-trained MobileFaceNet model (from

POSTERV2[12]), and the branch remains frozen in the initial stage of training. The SGD optimization algorithm was used to train the model, and the momentum parameter was configured as 0.9, the weight attenuation coefficient was 1×10^{-4}, and the initial learning rate was set as 1×10^{-3}. When the training process reached 60% and 80% of the total rounds, the learning rate was reduced by the attenuation factor of 0.1. The training batch size was set to 256, and all hyperparameters were filtered by the validation set grid search strategy. In the evaluation phase, mAP50 corresponding to the Intersection Over Union (IoU) threshold of 0.5 for RAF-DB dataset was used as the core index, and the training tasks were finally deployed on a server platform equipped with 6 RTX 6000 GPUs.

3.2 Experimental results

Table 1.Dataset RAF-DB comparison (AP50, \%).

Method	Year	Anger	Disgust	Fear	Happy	Neutral	Sad	Surprise	mAP50
SSD	2015	81.23	62.91	57.01	95.72	80.34	78.32	89.71	77.89
RetinaNet	2017	82.07	53.74	53.56	94.63	80.13	77.50	87.80	75.63
YOLOv3	2018	58.01	28.56	37.02	88.09	67.89	63.78	72.59	59.42
CenterNet	2019	53.26	17.41	30.62	91.32	75.36	66.83	84.63	59.92
EfficientNet	2019	68.72	52.25	45.47	93.75	78.67	76.96	84.31	71.45
YOLOv4	2020	39.25	0.00	10.36	87.61	52.77	45.72	59.91	42.23
YOLOv5	2020	45.75	8.86	0.00	91.77	64.73	65.60	74.36	50.15
YOLOX	2021	78.38	62.40	57.85	96.82	80.45	83.35	89.56	78.40
YOLOv7	2022	62.20	55.80	44.72	92.01	73.20	74.72	74.57	68.17
YOLOv8	2023	74.50	50.40	50.85	93.33	76.39	76.30	82.89	72.09
FER-YOLO-Mamba	2024	79.55	64.32	62.00	97.43	83.23	84.22	91.44	80.31
FER-YOLO-NCAMamba	2024	85.82	63.24	67.32	95.31	90.92	88.99	91.46	83.30
DEIM (Baseline)	2025	89.80	67.20	66.30	91.60	90.50	92.30	91.10	84.10
Natten-DEIM (Ours)	2025	94.28	72.73	66.59	95.72	90.84	88.16	89.62	85.28

As shown in Table 1, the proposed Natten-DEIM method achieves a 1.18 percentage points improvement in the mAP50 metric compared to the baseline model DEIM. From the category analysis, the performance of Anger and Disgust categories has achieved significant improvement, and the index is increased from 89.80% to 94.28% and from 67.20% to 72.73%, respectively. The Neutral category also increased from 90.50% to 90.84%; Happy category increased from 91.60% to 95.72%; The performance of Fear category remained stable. Sad and Surprise categories still maintain high scores as a whole. The experimental results show that the "channel recali-fication and scale alignment before fusion" operations performed by the NATTEN domain attention module and the multi-scale feature fusion module can effectively reduce the redundant information of cross-level features and strengthen the core features related to expression recognition.

Compared with the existing facial expression detection and recognition schemes, the performance advantages of Atten-DEIM are not only shown in the overall mAP50 evaluation index, but also improve the discrimination ability of individual emotions. Taking the YOLO improved model of FER-YOLO-Mamba and FER-YOLO-NCAMamba as an example, it achieves high detection accuracy on the RAF-DB dataset, but there is still a problem of low accuracy for Anger, Disgust and other categories. The NATTEN domain attention and multi-scale feature fusion module designed in this paper effectively suppresses the interference of cross-layer redundant information and significantly strengthens the feature expression ability of local regions closely related to expression discrimination by prefacing channel recali-fication and scale alignment operations in the feature fusion stage. The model achieves 4.48%and 5.53%increases in AP50 index on Anger and Disgust categories respectively compared with the baseline. For high-frequency emotion categories such as Happy and Neutral, the proposed method is superior to previous research in terms of stability and robustness while

maintaining or even improving the detection accuracy. On the premise of using a lightweight backbone network, Atten-DEIM achieves better performance than the existing expression detection methods, which fully verifies the application value of the proposed structure in actual deployment scenarios.

3.3 Model Visualization

(a) Model visualization results

(b) Model visualization heat map

Figure 4.Comparison of model visualization results(The sample facial images are from the Real-world Affective Faces Database (RAF-DB)[13]).

As shown in Figure 3, the left subfigure presents the feature localization output of Natten-DEIM, the delineation of the expression attention region in the form of a red box. The right subfigure is the attention heat map, which can intuitively see that the model can form a significant response enhancement to the key areas of expression discrimination such as eyes and mouth.

This visualization further verifies the performance benefits of Atten-DEIM: With the synergistic effect of domain attention and multi-scale feature fusion module, the model realized the ability upgrade from "rough localization of facial region level" to "accurate focus of key expression parts", which not only strengthened the local feature representation strongly related to expression classification, but also suppressed the redundant information response of non-relevant regions. It fully reflects its ability to accurately capture the core features of expression.

4. CONCLUSION

The stable geometric prior is introduced by LPB×EBB module, the controllable dual-gated modulation of channel and spatial dimension is realized by DFPM module, and the filtering process is completed before feature fusion by PCAF module, so as to reduce cross-layer redundancy and enhance feature complementarity. Experimental results show that the proposed method achieves the current optimal level on the RAFDB dataset, and the corresponding mAP50 index reaches 85.28%. Future research work will focus on improving the robustness and generalization performance of the model in complex real scenes.Although the Atten-DEIM model shows leading performance on the RAF-DB dataset, there are still many shortcomings. In the face of extreme occlusion, exaggerated expression, and high light/backlight complex scenes, the robustness of the model still needs to be improved. Its performance depends on the facial key point prior. If the key point detection fails or has a large deviation, it will significantly affect the overall performance of the model. The current research has only completed the verification on a single dataset, and the cross-dataset adaptation ability and cross-domain generalization performance of the model have not been systematically evaluated. Future research can be carried out in three directions: introducing more perfect occlusion modeling methods and invariance regularization strategies to enhance the adaptation ability of the model in complex scenes; The adaptive or unsupervised geometric prior update mechanism is explored to alleviate the dependence of the model on the accuracy of keypoint detection.

REFERENCES

[1] Linxiang Z ,Feifei L ,Jiawei C , et al. An improved feature pyramid network for object detection[J].Neurocomputing,2022,483127-139.DOI:10.1016/J.NEUCOM.2022.02.016.

[2] Liu S ,Qi L ,Qin H , et al. Path Aggregation Network for Instance Segmentation.[J].CoRR,2018,abs/1803.01534.

[3] Redmon J ,Divvala K S ,Girshick B R , et al. You Only Look Once: Unified, Real-Time Object Detection.[J].CoRR,2015,abs/1506.02640Ma X. M., Fu Y. P.*, Gao K. Z., Zhu L. H., Sadollah A., A multi-

objective scheduling and routing problem for home health care services via brain storm optimization. Complex System Modeling and Simulation, 2023, 3(1): 32-46.
[4] Arwidiyarti D. Single shot multibox detector (SSD) in object detection: a review[J]. IJACI: International Journal of Advanced Computing and Informatics, 2025, 1(2): 118-127.
[5] Tsung-Yi L ,Priya G ,Ross G , et al. Focal Loss for Dense Object Detection.[J].IEEE transactions on pattern analysis and machine intelligence,2020,42(2):318-327.DOI:10.1109/TPAMI.2018.2858826.
[6] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]//European conference on computer vision. Cham: Springer International Publishing, 2020: 213-229.
[7] Huang S, Lu Z, Cun X, et al. Deim: Detr with improved matching for fast convergence[C]//Proceedings of the Computer Vision and Pattern Recognition Conference. 2025: 15162-15171.
[8] Li F, Nie P, You M, et al. Triple-ATFME: triple-branch attention fusion network for micro-expression recognition[J]. Arabian Journal for Science and Engineering, 2025, 50(2): 807-823.
[9] Bin Talib H K, Xu K, Cao Y, et al. Micro-expression recognition using convolutional variational attention transformer (convat) with multihead attention mechanism[J]. IEEE Access, 2025.
[10] Deng J, Guo J, Xue N, et al. Arcface: Additive angular margin loss for deep face recognition[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 4690-4699.
[11] Bulat A, Tzimiropoulos G. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks)[C]//Proceedings of the IEEE international conference on computer vision. 2017: 1021-1030.
[12] Mao J ,Xu R ,Yin X , et al. POSTER++: A simpler and stronger facial expression recognition network[J].Pattern Recognition,2025,157110951-110951.DOI:10.1016/J.PATCOG.2024.110951.
[13] 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.

Dear reviewer,

Thank you very much for your comments and professional advice. These opinions help to improve academic rigor of our article. Based on your suggestion and request, we have made corrected modifications on the revised manuscript. Meanwhile, the manuscript had be reviewed and edited by language services of ELSEVIER. We hope that our work can be improved again. Furthermore, we would like to show the details as follows:

Reviewer 1#

1. In the introduction section, the definition of the research problem is vague, and it is suggested to illustrate the specific performance of the existing methods on these problems with specific experimental data or typical cases.

The author's answer: In the introduction part, the comparison of test results of different network models on the same data set is added.

2. The review of related work is not comprehensive, and it is recommended to cover the mainstream advanced methods in the field of facial expression recognition in recent years.

The author's answer: Add two newly published papers on facial expression recognition in 2025 to the introduction section.

3. The design details of the multi-head attention mechanism are missing, and it is suggested to clearly explain the value of the number of heads, the dimension setting of the query/key/value vector, and the initialization method of the output transformation matrix.

The author's answer: Add a structure diagram of the attention mechanism to the section on multi-head attention mechanism, and describe the relevant configuration information.

4. The paper should discuss the limitations of the model, including the challenges it may face when dealing with specific types of problems and potential directions for future improvements.

The author's answer: The dilemmas faced by the current model and potential future improvement directions are added to the final conclusion section.

5. The English expression of the paper is generally clear, but some sentences are somewhat lengthy. It is suggested to polish the language to improve the readability and professionalism of the paper.

The author's answer: Long sentences have been made as concise as possible.

Reviewer 2#

1. Please check the abbreviations of the whole manuscript, and provide the complete expression of the abbreviation when it FIRST appear, such as RAF-DB in the abstract, and those in the Introduction.

The author's answer: RAF-DB, which appears for the first time, has been revised to Real-world Affective Faces Database.

2. The end of the abstract should be summary of the significance of this work to the technical fields related to the conference scope.

The author's answer: The significance of the current research for the relevant technical fields has been added at the end of the abstract.

3. Figs. 1,3 use the photos of real people. The publisher may require copyright proofs and files to avoid the risk of law issue. If the relevant permission is not received, please replace the figures with the ones being permitted and copyrighted.

The author's answer: The real - person photos involved in this paper have cited the relevant papers of open - source datasets.

4. It is necessary to more logically present: how did the author design research methods based on theoretical principles and the research goals? How to select the parameters and indicators?

The author's answer: At the beginning of the section introducing the model methodology, an explanation has been added that the model is developed based on the structural properties of geometric priors combined with facial expression features, and designed to address the redundancy issue caused by multi-scale stacking, adopting the "modulation-first-then-fusion" design principle. In addition, model configuration parameters and evaluation metrics have been supplemented in the experimental setup section.

5. Discussion and results analysis need to more sufficiently summarize why this study and its results are significant compared to other related published works.

The author's answer: At the section of temporal result analysis, a comparative analysis of Average Precision (AP) across different classification tasks with previously published models has been added, demonstrating that the model designed in this paper is more statistically and practically meaningful.

6. References should be numbered according to the citation appearance order.However, ref.4-5 are cited before ref.1-3. Please correct all the numbers of references and citations.

The author's answer: The citation order of the references has been revised.

Yours sincerely,
Wei-Hong Luo,
January 7, 2025