Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 14, No. 3, 2025

Facial Expression Recognition with Hybrid Features Leveraging DINO Prior Knowledge

Yuansha Xie *, Cheng Ju, Yuxin Chang

Department of Data Science and Engineering, Yan'an University Xi'an Innovation College, Xi'an, Shaanxi, China
*** Corresponding author:** Yuansha Xie (Email: 1220767080@qq.com)

Abstract: Facial expression recognition plays a crucial role in smart education. To address the over-reliance on single prior image features or the ineffective integration of multiple image features in facial recognition tasks, as well as the poor generalization of facial expression recognition in natural environments. This study utilizes the large-scale visual model DINOv2 as a pre-training model, with its pre-trained weights frozen, leveraging its learned experience from natural image datasets to acquire more universal image features, thereby enhancing the generalization performance of feature extraction. Furthermore, this work proposes a Hybrid Feature Facial Expression Recognition model (HFFER). The model utilizes two different pre-trained models to acquire distinct features, and effectively integrates them through cross-attention mechanisms and multiple convolutions. Experimental results demonstrate that the model achieved accuracies of 92.18% on the RAF-DB datasets, respectively, surpassing or being comparable to existing models. This study introduces a novel approach to facial expression recognition, while its application in real classroom images demonstrates its feasibility and potential in practical educational settings.

Keywords: Facial Expression Recognition; The Large-scale Visual Model DINOv2; Hybrid Feature Facial Expression Recognition Model.

1. Introduction

Facial Expression Recognition (FER), as a downstream task in the field of computer vision, plays a significant role in identifying and understanding human facial expressions. It has wide-ranging applications in human-computer interaction, smart education, emotion analysis, and virtual reality. In recent years, the FER domain has transitioned from traditional methods to deep learning techniques. Early studies employed manually selected features such as Oriented Histogram of Gradients and grayscale textures for analyzing facial expressions, while later approaches utilized architectures like Convolutional Neural Networks [1] and Transformers [2] to automatically learn features for classification. Most existing work builds upon established network architectures, modifying them for FER tasks. For instance, dilated convolution modules or anti-aliasing modules are introduced into Convolutional Neural Networks, variational learning with knowledge modulation is applied to VAE (Variational Autoencoder) architectures to enhance recognition of novel expression types [3], or dual-stream feature extraction mechanisms are incorporated into Transformer architectures [4].These methods have made good progress in the task of facial expression recognition. However, these works highly rely on the prior facial features provided by pre trained weights obtained from a large number of annotated facial image datasets, which results in the features learned by the model not fully reflecting the complexity of facial expressions or the diversity of background textures, leading to overfitting and limiting the performance of the model. Recently, the facial expression recognition model POSTER proposed by Zheng et al[4]. effectively alleviated the above-mentioned problems. Specifically, this work introduces additional facial landmarks as feature priors to guide the extraction of facial expression features, and improves the overall performance of the model by extracting image features from different perspectives and fusing them. However, this method leads to the accuracy of facial expression recognition being affected not only by the backbone network of facial feature extraction, but also by the accuracy of facial keypoint algorithms, thereby introducing more unstable factors. This has prompted our thinking: can we effectively improve the accuracy of expression recognition models by utilizing the high generalization features contained in general visual models (such as DINOv2) without relying on more complex labeled facial image data?

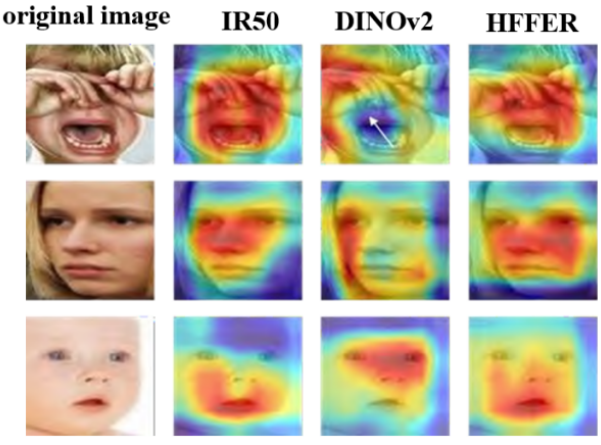

Fig 1. Feature Visualization Illustrations of IR50 and DINOv2

Inspired by the transfer learning method of pre trained models [5-7], this model introduces the DINOv2 [8] visual big model to extract image generalization features. DINOv2 utilizes a large pre training set of high-quality natural images and effective self supervised training methods to learn general features from the images on its own. Compared with common facial recognition networks, DINOv2 is more sensitive to capturing global information. To validate this idea, we

visualized the features obtained from two models: the facial recognition network (IR50 [9]) and DINOv2. The darker the red area in the figure, the higher the level of attention. From Figure 1, it can be observed that IR50 performs outstandingly in capturing facial detail features, while DINOv2 is better at extracting the overall features of the image. For example, in the first row, IR50 mainly focuses on the overall facial area due to training samples, while DINOv2 pays more attention to wiping tears and hand movements that contain more semantic information (as indicated by the arrow area). In addition, when encountering side face images, the IR50's ability to capture and locate facial detail features is weakened, exposing some shortcomings of poor generalization, while DINOv2 can still capture contour information normally.

Based on the above observations, we believe that if the complementarity of the two backbone networks can be fused and fully utilized, it can improve the accuracy of feature capture and adaptability to complex samples, thereby enhancing generalization ability. In this article, unlike the common one-way attention fusion operation in feature fusion work, we propose a bidirectional attention fusion mechanism. By using facial detail features from IR50 and image generalization features from DINOv2 as joint references to facilitate attention fusion operations, it avoids the potential for one-way fusion mechanisms to overly rely on the attention area provided by a certain type of feature after optimization, thereby avoiding perceptual bias towards facial features and preventing the problem of difficulty in capturing salient areas in optimized facial features.

The main contributions of this article include:

A novel facial expression recognition framework HFFER has been proposed, and effective exploration of visual large-scale models in facial expression recognition tasks has been carried out, verifying the complementarity between general visual learning frameworks and classical facial expression recognition models. A hybrid feature fusion module based on bidirectional attention is proposed, which fully utilizes the rich understanding of images by the pre trained visual model DINOv2 and the targeted feature learning improvement of facial images by traditional facial recognition models. By relying on the complementarity of different source clues, the facial expression recognition performance is effectively improved. The proposed method has achieved competitive results on the commonly used public facial expression recognition dataset RAF-DB.

2. Related Work

2.1. Facial Expression Recognition on Headings

Early facial expression recognition methods relied on manually selecting features. Zhao et al. [10] used LBP (Local Binary Pattern) to study FER and achieved good results. However, the selection of these features is usually based on the knowledge and experience of domain experts, rather than automatically learned by algorithms. Therefore, once these methods are applied to specific scenarios outside the laboratory, the usability of the algorithm is greatly reduced. The development of convolutional neural networks has led people to explore the usability of convolutional neural networks in facial expression recognition tasks. Savchenko et al. [11] studied lightweight convolutional neural networks such as MobileNet, Efficientnet, RexNet for FER task learning and verified the effectiveness of CNN for FER.Wang Jun et al. Lan Zhengjie et al. [12] enhanced the expressive power of the model by combining word frequency inverse document frequency and spatial pyramid attention mechanism to fuse features. VTFF [13] performs well in processing outdoor facial expression recognition tasks by fusing LBP features and convolutional features of images. In addition, facial keypoint detection technology, as a technique for identifying and estimating the positions of predetermined keypoints such as eyebrows, eyes, and mouth on the face, provides an important foundation for accurate recognition of facial expressions. Considering the rich features contained in facial keypoints, researchers have adopted different methods to utilize this information. Nakamura et al. [14] used facial keypoint features as constraints for the latent distribution input into VAE (Variational Autoencoder), effectively alleviating overfitting problems. After considering the relationship between facial keypoint features and image features, Zheng et al. proposed POSTER [4], which achieved state-of-the-art performance by cross fusing image features with facial landmark features. Inspired by the POSTER model, this article also adopts a feature cross fusion strategy to enhance the model's generalization ability. In the process of extracting the main features of the image, this study chose to use IR50 as the backbone network. Unlike the POSTER model, this paper does not use facial keypoint features, but instead chooses the large visual model DINOv2 trained on natural images to extract image features, in order to learn more universal image features.

2.2. Visual Large Model

With the revolutionary success of Transformer in the field of natural language processing (NLP), people are beginning to try to apply it to the visual domain. As the first model to apply Transformer to the field of vision, ViT (Vanilla Vision Transformer)'s successful exploration has officially entered the era of Transformer for visual models. DINO [15] is a large model framework for self supervised learning in the field of visual understanding. With the collaboration of ViT, it adopts an unlabeled knowledge distillation method, which can demonstrate superior image semantic expression and resolution ability than unsupervised learning without inputting human understanding information. In addition, He et al. [6] proposed a masked autoencoder (MAE) applied in the field of computer vision, which demonstrates powerful image reconstruction and transfer learning capabilities as an extensible self supervised learning method. MetaAI subsequently introduced a universal image segmentation model called SAM, which can return an effective segmentation mask upon receiving any given segmentation prompt, enabling downstream segmentation tasks without fine-tuning and achieving powerful zero sample generalization, greatly promoting the development of computer vision infrastructure models. However, SAM's performance in common visual tasks other than semantic segmentation was not satisfactory. To fill this gap, Meta introduced DINOv2. While maintaining the advantages of DINO, the performance has been further improved, showing excellent performance in downstream visual tasks such as image classification, semantic segmentation, and depth estimation.

3. Facial Expression Recognition Model based on Hybrid Feature Network

3.1. Overall Architecture of the Model

The overall architecture of facial expression hybrid feature recognition based on DINOv2 is shown in Figure 2. To investigate the effectiveness of pre trained image feature extraction models in downstream tasks of facial expression recognition, this paper proposes a novel facial expression recognition model HFFER. The HFFER model mainly consists of three parts: preliminary feature extraction, mixed feature fusion, and classification prediction. In the initial feature extraction stage, the model utilizes IR50 and DINOv2 to capture feature maps of facial image data. By utilizing the complementarity of feature representation between the two, the model can improve its understanding and recognition performance of facial expressions.

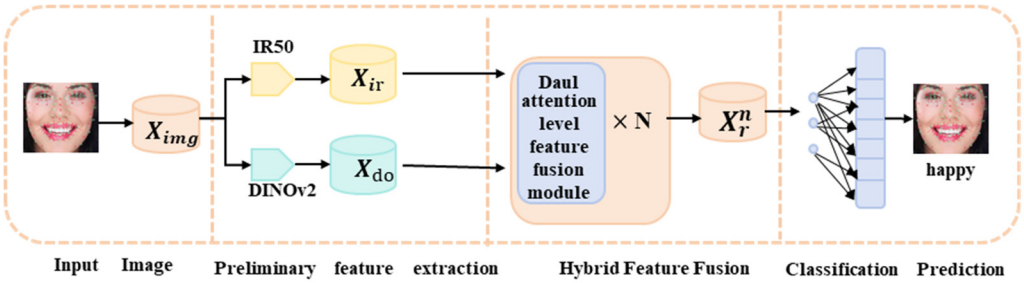

Fig 2. Overall Structure of the Proposed Model HFFER

In the mixed feature fusion stage, this work aims to improve the performance of facial expression recognition by utilizing the complementarity of two feature types. Unlike traditional one-way attention fusion mechanisms , we introduce a dual attention level feature fusion module. Specifically, the HFFE module aims to solve the fusion problem between low-level features (including noise but rich in details) and high-level features (semantic abstraction but loss of details) in traditional U-Net encoders, by enhancing feature representation through cross layer attention mechanisms and suppressing background interference. HFFE enables efficient interaction between high-level and low-level encoder features through a four-step procedure: dual-branch input alignment, spatial attention refinement, dynamic weight calibration, and coordinate attention fusion. The implementation operates as follows:

Input Features: Adjacent encoder features from two layers are received—the low-level feature X_{ir} from the previous layer (rich in detail but noise-sensitive) and the high-level feature X_{do} from the current layer (semantically abstract with lower resolution).

Spatial Attention (SAM) Processing: SAM is applied separately to the aligned X_{ir} and X_{do} to highlight target regions, yielding refined features X'_{ir} and X'_{do}.

Weight Matrix Generation: A 1×1 convolution (CBR1) followed by a sigmoid function computes feature importance weights.

Feature Calibration: The weight matrix is applied to the original features via element-wise multiplication to suppress noisy areas.

CoordAtt Fusion: The calibrated X''_{ir} and X''_{ir} are concatenated, and spatial position information is encoded through CoordAtt to produce a fusion weight matrix $\mathrm{SWM}_{\mathrm{fuse}}$.

Hierarchical Encoding Output: The final HFFE output X_{HFFE} is obtained by fusing the calibrated features with $\mathrm{SWM}_{\mathrm{fuse}}$ via element-wise multiplication.

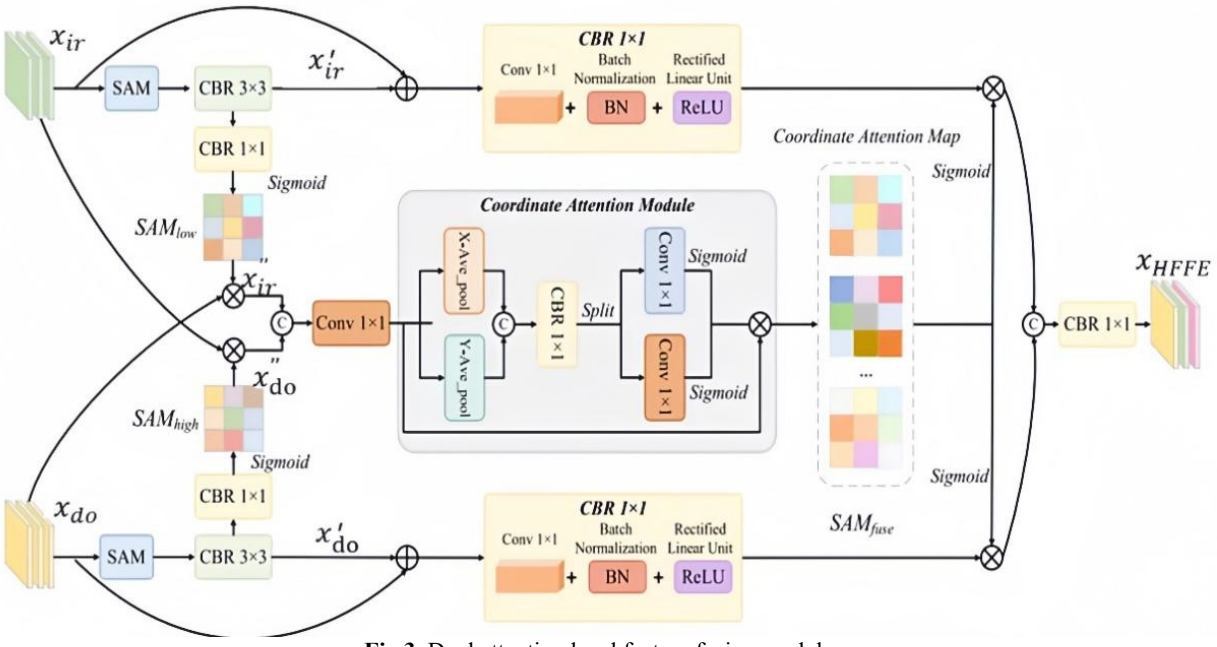

Fig 3. Dual attention level feature fusion module

3.2. Preliminary Feature Extraction of Facial Expressions

The feature extraction layer aims to extract diverse feature maps from input images through two deep backbone networks pre trained on different datasets, IR50 and DINOv2, and comprehensively utilize the advantages of both models to perceive facial feature clues more comprehensively and richly from different perspectives. Specifically, IR50 focuses on extracting detailed facial features such as overall facial contours, forehead wrinkles, etc. In contrast, the latter can capture more general image features, thereby enriching the representation of facial information. It is worth mentioning that in our work, we froze the pre training weights of DINOv2 and only optimized the IR50 model and subsequent fusion prediction modules, ensuring that the high generalization ability of DINOv2 models trained on a large number of real-world corpora is not compromised. Our model inputs facial expression images $X_{img} \in RH \times W \times 3$ into the backbone networks IR50 and DINOv2 respectively, obtaining deep facial detail features $X_{ir} \in RH' \times W' \times D$ and image generalization features $X_{img} \in RH' \times W' \times D$ extracted under the influence of different data priors and network structures. Among them, H and W respectively represent the height and width of the input image, H' and W' represent the height and width of the output feature map, and D is the feature dimension. During the training process, we only fine tuned and optimized the backbone network IR50, while freezing the parameters of the DINOv2 model.

3.3. Hybrid Feature Fusion Module Based on Bidirectional Attention

This work focuses on the asymmetric limitations of the unidirectional attention mechanism fusion mechanism in handling feature fusion. The features fused through a one-way attention mechanism often overly rely on the attention area provided by a certain type of feature, which may lead to perceptual bias towards facial features and affect the effectiveness of feature fusion. In the unidirectional attention mechanism, one feature is selected as the primary feature, while another feature assists in determining the attention area. The bidirectional attention mechanism comprehensively captures key information by jointly referencing two features, as shown in Figure 3.

Given the asymmetric limitations of the unidirectional fusion mechanism and inspired by the Transformer fusion model [16], we propose a bidirectional attention fusion module suitable for mixed features. For each bidirectional attention fusion module, its input features consist of three parts, namely facial detail features X_{ir}, image generalization features X_{do}, and fusion features output from the previous layer X_{n-1}. Considering the strong correlation between IR50 and facial expression encoding, we define X_0 as the initial state of X_{ir} as the fusion feature X_f.

Specifically, the workflow of the first fusion module can be simplified as X_{ir}'s self attention learning and $\{X, X\}$'s cross attention fusion, but this simplification can only represent a special case. In fact, starting from the second module, our fusion operation can be expressed as a cross attention fusion process of $\{X_{n-1}, X_{ir}\}$ and $\{X_{n-1}, X_{do}\}$ feature pairs.

Traditional convolutional operations are inherently constrained by their fixed receptive field, limiting their capacity to capture long-range dependencies. Standard 3 × 3 or 5 × 5 convolutions are primarily effective for extracting local spatial information, yet they struggle to distinguish small infrared targets from complex backgrounds effectively. Noise interference and variations in target contrast further exacerbate this issue, posing significant challenges to robust small target detection. In addition, infrared small targets exhibit diverse scale distributions across different scenes, rendering single-scale feature extraction methods inadequate for capturing multiscale target variations. As shown in Figure 3, HFFE takes encoding features from the previous level X_{ir} and the current level X_{do} as inputs. Given the resolution discrepancy between these feature maps, X_{do} undergoes bilinear upsampling to match the spatial dimensions of X_{ir}. Each feature map is then independently processed using SAMs to highlight target-relevant regions, producing refined feature representations.

$$X'_{\mathrm{ir}} = CBR_3\left(S_{att}(X_{\mathrm{ir}})\right) \tag{1}$$

$$X'_{\mathrm{do}} = CBR_3\left(S_{\mathrm{att}}\left(\mathcal{B}i(X_{\mathrm{do}})\right)\right) \tag{2}$$

Where S_{att} represents the spatial attention operation.

To ensure effective cross layer feature interaction, HFFE generates spatial weight matrices (SWMs), which adaptively recalibrate feature importance across different scales. These matrices are computed using channel-wise recalibration and are formulated as follows:

$$SWM_{ir} = \sigma\left(CBR_1(X'_{ir})\right) \tag{3}$$

$$SWM_{do} = \sigma\left(CBR_1(X'_{do})\right) \tag{4}$$

where SWM represents the generated spatial weight matrix. The recalibrated feature maps are then obtained through element-wise multiplication with their respective SWMs

$$X''_{\mathrm{ir}} = X_{\mathrm{ir}} \otimes SWM_{\mathrm{ir}} \tag{5}$$

$$X''_{\mathrm{do}} = X_{\mathrm{do}} \otimes SWM_{\mathrm{do}} \tag{6}$$

To further enhance the synergy between different features,the coordinate attention (CoordAtt) is incorporated, which encodes spatial dependencies along both the horizontal and vertical directions through a global pooling operation. Unlike conventional channel attention mechanisms, CoordAtt retains positional information crucial for small target detection. The fused coordinate attention weights are generated as

$$\mathrm{SWM}_{fuse} = \sigma\left(\mathrm{CA}(\mathrm{Conv}_{1\times1}[X'_{\mathrm{ir}}, X'_{\mathrm{do}}])\right) \tag{7}$$

The final hierarchical encoding feature FHFFE is then computed as

$$X_{HFFE} = \mathrm{Conv}_{1\times1}\begin{bmatrix} SWM_{fuse} \otimes \left(\mathrm{CBR}_1(X'_{\mathrm{ir}} + X_{\mathrm{ir}})\right) \\ SWM_{fuse} \otimes \left(CBR_1(X'_{\mathrm{do}} + X_{\mathrm{do}})\right) \end{bmatrix} \tag{8}$$

By integrating HFFE into the network, the model effectively leverages the statistical differences among multilevel encoder features, ensuring robust feature aggregation and noise suppression. Compared to traditional skip connections, HFFE refines feature interactions at multiple scales, reducing false positives caused by DSPM while enhancing the discriminative capacity of hierarchical encoding features. The refined hierarchical representations generated by HFFE are subsequently fed into the decoder stage through HSC, where they further enhance target localization and improve detection accuracy in complex infrared scenes.

3.4. Loss Function

To measure the predictive ability of the model, a supervised training approach is adopted. This article did not use a complex loss function calculation method, but instead used a combination of Label Smooth Cross Entropy Loss and Cross Entropy Loss to optimize the loss, as shown in formula (9).

$$Loss = 0.67 \times L_{lsce} + 0.33 \times L_{ce} \tag{9}$$

Among them, L_{lsce} is the label smoothing cross entropy loss function of the sample, and the specific calculation method is shown in formula (10). L_{ce} is the standard cross entropy loss function of the sample, and the specific calculation method is shown in formula (11).

$$L_{lsce} = -\frac{1}{N}\sum_{i=1}^{N}\left((1-\grave{o})\ln(p_i) + \frac{\grave{o}}{C}\sum_{j=1}^{c}\ln(q_{ij})\right) \tag{10}$$

$$L_{ce} = -\frac{1}{N}\sum_{i=1}^{N}\left(p_i \ln\left(\frac{e^{q_i}}{\sum_{j=1}^{c} e^{q_{ij}}}\right)\right) \tag{11}$$

Among them, N is the number of samples, C is the number of categories, p_i is the true label probability value of sample i, q_{ij} is the probability that the model predicts sample i to be in category j, $\grave{o}$ is the smoothing parameter, usually taking a smaller value. This model is set to 0.2 by default. q_i is the probability of the sample i being predicted by the model as belonging to each category.

4. Experimental Results and Analysis

This experiment validated the effectiveness of the model and explored the impact of each module on the overall performance of the model on two standard facial expression recognition task (FER) datasets, including RAF-DB [17]. Align image data, with a training set of 12271 images and a testing set of 3068 images.

Fig 4. Dataset Example

Facial images related to keywords mainly revolve around basic human emotions, including neutral, happy, angry, sad, fearful, surprised, disgusted, and contemptuous. This experiment selected RAF-DBt (7 classes) based on 7 emotion categories (excluding contempt) for validation.

Table 1. Statistics of Datasets

dataset	training set	test set	number of categories
RAF-DB	12271	3068	7

The experiment utilized two NVIDIA RTX 3090 GPUs under the PyTorch framework for end-to-end training and inference of the model. The model selects the IR50 backbone network pre trained on the Ms-Celeb-1M dataset [to extract facial detail features, and updates the weights of the image backbone during the training process. Under the premise of freezing weights, select the Base version in pre trained DINOv2 to obtain the generalized features of facial expression images. The input image size parameter is H, W=224, and the output feature map size parameter H',

W'=7, with a feature dimension of 768. The number of stacked bidirectional feature fusion modules in the model is 8. During training, the batch size is set to 200 and optimized using Adam optimizer with an initial learning rate of 4×10^{-5}. Exponential LR is used for learning rate scheduling, and the learning rate decay factor γ is set to 0.99.

In this work, we utilized publicly available data RAF-DB to test and validate the effectiveness of the proposed model. The comparison results between the HFFER model proposed in this article and other advanced algorithm models on the RAF-DB dataset are shown in Table 2. To ensure a fair comparison environment, we retrained all open-source works according to the parameters in the article. The experimental results on the RAF-DB dataset show that the proposed model achieves the best performance, with an accuracy of 92.18%, which is higher than other excellent facial expression recognition models. Compared with TransFER [18] and POSTER [4], which perform well on this dataset, it has improved by 1.27 and 0.13 percentage points, respectively. Compared to other advanced facial expression recognition methods, the HFFER model has achieved better or comparable performance in seven categories of datasets. Our work has achieved results similar to POSTER and superior to other advanced works without relying on facial keypoint annotation.

Table 2. Comparison of Experiment Results of Models

dataset	Release Year	Facial key points	RAF-DB
LDL-ALSG[19]	CVPR2020	×	85.53
DACL[20]	WACV2020	×	87.78
KTN[21]	TIP 2021	×	88.07
DMUE[22]	CVPR 2021	×	89.42
Meta-Face2Exp[23]	CVPR2022	×	88.54
EAC[24]	ECCV2022	×	90.35
POSTER[4]	ICCVW2023	√	92.05
HFFER	—	×	92.18

5. Summary

In this study, we propose a hybrid feature bidirectional attention fusion network based on DINOv2 prior for facial expression recognition tasks. This model combines facial detail features and image generalization features extracted by two pre trained models, IR50 and DINOv2, through an effective cross fusion mechanism. This work apply the visual macro model DINOv2 to facial expression recognition tasks under the premise of frozen weights and demonstrate its gain effect on the task. This discovery is expected to provide new directions for the development of this field. Our experimental results on the RAF-DB public dataset have validated the feasibility of our design motivation and the rationality of our design scheme. In addition, we further tested the practical application of facial expression recognition on real classroom images, and the application results demonstrated the feasibility of the proposed model in actual educational scenarios, demonstrating the potential of this work to support sentiment analysis and learning status monitoring in smart education. In the future, we will conduct research on model lightweighting, aiming to further reduce the model size while maintaining performance, and improve its operational efficiency on small computing devices, so that our model can adapt to more teaching scenarios such as outdoor teaching, physical education courses, etc. At the same time, we will actively explore the joint learning of natural language prompts and visual information, in order to achieve more accurate and intelligent cross modal understanding and application.

Acknowledgments

The authors gratefully acknowledge the financial support from the Scientific Research Prolectsot Yan' an University Xi an Innovatior College, including grant number 2025XJKY09.

References

[1] HE K M, ZHANG X Y, REN S Q, et al. deep residual learning for image recognition[C]‖Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Washington D.C., USA: IEEE Press, 2016: 770 778.

[2] VASWANI A, SHAZEER N M, PARMAR N, et al. Attention is all you need[C]‖Proceedings of the 31st International Conference on Neural Information Processing Systems. New York, USA: ACM Press, 2017: 6000-6010.

[3] ZHU J, LUO B, YANG T, et al. Knowledge conditioned variational learning for one-class facial expression recognition [J]. IEEE Transactions on Image Processing, 2023, 32: 4010-4023.

[4] ZHENG C, MATIAS M, CHEN C. POSTER: a pyramid cross-fusion transformer network for facial expression recognition[C] ‖Proceedings of IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Washington D.C., USA: IEEE Press, 2022: 3138-3147.

[5] KIRILLOV A, MINTUN E, RAVI N, et al. Segment anything[C]‖Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Washington D.C., USA: IEEE Press, 2023: 3992-4003.

[6] HE K M, CHEN X L, XIE S N, et al. Masked autoencoders are scalable vision learners [C]‖Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D.C., USA: IEEE Press, 2022: 15979-15988.

[7] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale[C]‖Proceedings of the 9th International Conference on Learning Representations (ICLR). [S.l.]: AAAI Press, 2021: 12-18.

[8] OQUAB M, DARCET T, MOUTAKANNI T, et al. DINOv2: learning robust visual features without supervision [EB/OL]. [2024-02-05]. https:‖openreview.net/forum? id=a68SUt6zFt.

[9] DENG J K, GUO J, YANG J, et al. ArcFace: additive angular margin loss for deep face recognition [C]‖ Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D.C., USA: IEEE Press, 2018: 4685-4694.

[10] ZHAO G Y, PIETIKAINEN N. Dynamic texture recognition using local binary patterns with an application to facial expressions [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2007, 29(6): 915-928.

[11] SAVCHENKO A V. Facial expression and attributes recognition based on multi-task learning of lightweight neural networks[C]‖Proceedings of the 19th InternationaSymposium on Intelligent Systems and Informatics (SISY). Washington D.C., USA: IEEE Press, 2021: 119-124.

[12] LAN Z J, WANG L,NIE X. An expression recognition algorithm based on term frequency-inverse document frequency and hybrid loss[J]. Computer Engineering,2023, 49 (1): 295-302, 310.

[13] MA F Y, SUN B, LI S T. Facial expression recognition with visual transformers and attentional selective fusion[J]. IEEE Transactions on Affective Computing, 2021, 14: 1236-1248.

[14] NAKAMURA F, MURAKAMI M, SUZUKI K, et al. Analyzing the effect of diverse gaze and head direction on facial expression recognition with photo-reflective sensors embedded in a head-mounted display[J]. IEEE Transactions on Visualization and Computer Graphics, 2023, 29(10): 4124-4139.

[15] CARON M, TOUVRON H, MISRA I, et al. Emerging properties in self-supervised vision transformers [C]‖ Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Washington D.C., USA: IEEE Press, 2021: 9630-9640.

[16] XIE B, LIU Y Q, LI Y L. Colorectal polyp segmentation method combining polarized self-attention and Transformer [J]. Opto-Electronic Engineering, 2024, 51(10): 240179.

[17] LI S, DENG W H, DU J P, et al. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild [C]‖Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Washington D.C., USA: IEEE Press, 2017: 2584-2593.

[18] RAN R S, WENG W W, WANG N, et al. Expression recognition based on the extraction of key facial features[J]. Computer Engineering,2023,49(2):254-262.

[19] CHEN S K, WANG J F, CHEN Y D, et al. Label distribution learning on auxiliary label space graphs for facial expression recognition [C]‖Proceeding of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D.C., USA: IEEE Press, 2020: 13981-13990. FARZANEH A H, QI X J. Facial expression recognition in the wild via deep attentive center loss [C]‖Proceedings of 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). Washington D.C., USA: IEEE Press, 2021: 2401-2410.

[20] LI H Y, WANG N N, DING X P, et al. Adaptively learning facial expression representation via C-F labels and distillation [J]. IEEE Transactions on Image Processing, 2021, 30: 2016-2028.

[21] SHE J H, HU Y B, SHI H L, et al. Dive into ambiguity: latent distribution mining and pairwise uncertainty estimation for facial expression recognition [C]‖Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D.C., USA: IEEE Press, 2021: 6244-6253.

[22] ZENG D, LIN Z K, YAN X, et al. Face2Exp: combating data biases for facial expression recognition [C]‖Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D.C., USA: IEEE Press, 2022: 20259-20268.

[23] ZHANG Y, WANG C, LING X, et al. Learn from all: erasing attention consistency for noisy label facial expression recognition[EB/OL].[2024-02-05].https:‖arxiv.org/pdf/ 2207. 10299.