Face Expression Recognition via Product-Cross Dual Attention and Neutral-Aware Anchor Loss

Yongwei Nie[1][0000-0002-8922-3205], Rong Pan[1], Qing Zhang[2], Xuemiao Xu[1], Guiqing Li[1], and Hongmin Cai ✉[1]

[1] South University of Technology, Guangzhou, 510006, China
✉Corresponding author: hmcai@scut.edu.cn
[2] Sun Yat-sen University, Guangzhou, 510006, China

Abstract. Face expression recognition is an important task whose aim is to classify a face image to a kind of expression such as happy, sad, or surprise, etc. This task is challenging due to the ambiguities in expressions and also in the diverse poses and occlusions of the head. To handle this challenging task, recent approaches usually rely on attention mechanism to make the network focus on the most critical regions of a face, or apply a consistency loss that enforces extracting similar features from the same expressions. This paper proposes a new attention mechanism that combines the advantages of dot-product attention and feature cross-attention. The proposed new product-cross dual attention mechanism can better leverage the landmarks to extract more discriminative features from an input image. Second, although previous approaches can enforce similarity between features of the same expressions, they do not consider the arousal degree of an expression. We propose a neutral-expression-aware expression feature similarity loss based on the traditional anchor loss, which can further guide the network to learn better features from an input image. Extensive experiments demonstrate the advantages of our method over previous approaches.

Keywords: Face expression recognition · attention mechanism · expression arousal degree · face landmark.

1 Introduction

Facial expressions are one of the most powerful, natural, and universal signals that humans use to convey emotional states and intentions [8]. Psychologists Ekman and Friesen [13] proposed that human emotions can be expressed through six basic expressions: surprise, sadness, disgust, happiness, fear, and anger (neutral expressions have also been included in recent years). As a fundamental task in computer vision, facial expression recognition, i.e., recognizing the kind of expression in a face image, has great applications in many image and video analysis tasks [6, 7, 23, 28, 33, 34, 36, 49, 54].

Traditional FER methods typically rely on manual feature extraction or shallow learning techniques such as Local Binary Pattern (LBP) [41], Non-negative

Matrix Factorization (NMF) [58], and Sparse Learning [59]. Shan introduced Local Binary Pattern to describe local texture features in images, David proposed the Scale-Invariant Feature Transform (SIFT) [26] to enhance tolerance to noise, lighting, and other interferences. In 2008, Bashyal [2] presented a method for extracting expression features based on Gabor wavelet transform. In recent years, an increasing number of approaches have shifted towards deep learning techniques, including convolutional neural networks (CNNs) [20], generative adversarial networks (GANs) [15], transformers [44], for the extraction and classification of facial expression features. These methods have achieved state-of-the-art recognition accuracy, significantly surpassing results obtained by traditional machine learning methods.

Despite the increasing number of methods aimed at improving the accuracy of facial expression recognition, they continue to grapple with the inherent challenges of this task: 1) Intra-class Variability: the same emotion can manifest with significant variations in facial shape and intensity across different faces. 2) Inter-class Similarity: different individuals may share similar features even among distinct facial expressions (e.g., in regions like the forehead and cheeks).

To address the aforementioned issues, many different approaches have been developed. Some studies leverage auxiliary tasks related to facial expression recognition to enhance accuracy. For example, Chang et al. [4] summarized AU labeling rules from FACS, then designed facial partitioning schemes to extract local facial region features using a backbone, and finally used the correlations between features from different regions to guide the training of the feature learning framework. Li et al. [25] employed AU recognition as an auxiliary task, facilitating mutual improvement between the two tasks by summarizing the distribution relationship between expressions and AUs. Recently, Xue et al. [50] introduced a Transformer-based approach called TransFER. After extracting feature maps using a backbone CNN, they designed local CNN blocks to pinpoint different local patches, and subsequently, used a Transformer encoder equipped with multi-head self-attention modules to compute global relationships among these local patches. Zheng et al. [57] proposed POSTER, which utilizes a pyramid cross-fusion Transformer to explore the correlation between image features and landmark features, aiming to address issues related to inter-class similarity, intra-class variation, and scale sensitivity in facial expression recognition.

Summarizing the advantages of all the above approaches, we find that, to achieve higher accuracy, the network needs to focus on the most critical facial regions, and to our best knowledge, the above recent advances [50, 57] achieve this by using attention mechanism such as that in Transformer [44]. Inspired by the POSTER approach [57], this paper proposes a combined dot-product and feature-cross dual attention mechanism. Specifically, we follow POSTER to employ a landmark detector to extract landmark position features and utilize these features to calculate an attention map. Our dot-product attention is then implemented as the multiplication between the attention map and the image features extracted by the backbone, which guides the network to filter irrelevant features away from the crucial facial areas. After that, we further feed the attended features to a cross-attention module, which uses the landmark feature as the query, and the image feature as the key and value to update the image features. The sequentially applied two kinds of attention mechanisms constrain the network to focus on specific facial regions, reducing the importance of irrelevant areas and minimizing their impact on facial expression recognition.

Besides the network's own capability in identifying discriminative features by attention mechanism, many researchers also explicitly address issues related to intra-class variation and inter-class similarity by applying loss functions such as center loss [48], anchor loss [40], and locality-preserving loss [24] etc. These losses aim to increase inter-class distances or decrease intra-class distances, thereby mitigating intra-class variation and inter-class similarity problems in facial expressions. However, these losses treat all expressions equally and overlook the differences in intensity that exist among different expressions. For example, within happy expressions, there can be significant dynamic differences between a smile and a hearty laugh, whereas neutral expressions typically lack such dynamic variations. Additionally, if an expression's intensity is relatively weak, the network may easily misclassify it as a neutral expression. Consequently, neutral expressions require distinct treatment. The second contribution of this work is that we enhance the anchor loss to accommodate the characteristics of neutral expressions and propose the so-called neutral-expression-aware anchor loss. It strengthens the ability to distinguish between neutral expressions and other expressions by constraining the features of neutral expressions.

In summary, the contributions of this paper are

- We propose a product-cross dual attention mechanism. On one hand, we incorporate landmarks into the computation by taking their product with facial expression features to adjust the weights of facial regions. On the other hand, within the ViT [11] architecture, we calculate cross-attention between landmark features and image features to reinforce the network's focus on crucial areas.
- We propose a neutral-expression-aware anchor loss, which improves the original anchor loss with the characteristics of neutral expressions to strengthen the network's ability to distinguish between neutral expressions and other expressions.
- Experimental results on several datasets demonstrate that our approach yields superior performance compared to other methods.

2 Related Work

2.1 Landmark

The auxiliary tasks of facial expression recognition typically include facial attribute prediction, facial landmark detection, facial recognition, and facial action unit detection, among others. Many approaches [19, 25, 30, 37, 53] improve facial expression recognition accuracy by jointly training multiple auxiliary tasks. Among these auxiliary tasks, facial landmark detection has matured significantly.

For expression recognition, facial landmark detection provides valuable facial geometry information, and when combined with spatial image features, it effectively enhances the accuracy of expression recognition. Additionally, facial landmarks accurately locate key facial regions such as eyes, mouth, eyebrows, which are crucial for expressing emotions. By pinpointing these locations, the network can narrow down its focus on these areas when selecting facial features. Jung and colleagues [18] employed two deep networks: the first one extracts temporal appearance features from images, while the second one extracts temporal ensemble features from facial landmarks. They used a novel fusion method to combine these two models, resulting in improved performance in expression recognition. In the case of POSTER [57], they utilized a pre-trained facial landmark detection model, MobileFaceNet [5], to extract landmark features. They designed a pyramid structure and a dual-stream structure, and calculated cross-attention between the features extracted from the backbone and the landmarks. In this study, pre-trained landmark detectors were used to extract geometric features and compute attention maps, guiding the network's focus towards crucial facial areas and reducing the weight assigned to irrelevant regions.

2.2 Transformer in FER

The powerful attention mechanism within the Transformer architecture [44] has led to leading results in various computer vision domains. In the field of facial expression recognition, it is common to cascade Convolutional Neural Networks (CNNs) with Transformers, feeding the features extracted by CNNs into Transformers for attention computation. Ma et al. [27] were among the first to introduce Transformers into facial expression recognition. They extracted features from RGB and LBP images and used an Attention Selective Fusion module (ASF) to merge global and local features. These merged features were then transformed into visual tokens and fed into a Multi-layer Transformer for encoding. Xue et al. [50] incorporated a multi-attention dropping module within the Transformer, enabling the model to extract comprehensive local information from every part of the face, rather than just the most discriminative parts. Additionally, they designed a pooling module within the Transformer to progressively reduce the number of tokens in the blocks, eliminating information irrelevant to expression recognition. Zheng et al. [57] employed cross-attention within the Transformer, exchanging Query matrices between the image stream and landmark stream, facilitating the fusion of features from both streams.

2.3 Losses used in FER

Due to the characteristics of expressions having intra-class variations and inter-class similarities, various loss functions have been applied in networks to increase inter-class distances and reduce intra-class distances. Ruan et al. [39] employed a compactness loss to learn class centers, aiming to ensure that features from different images of the same expression are close to the respective class centers.

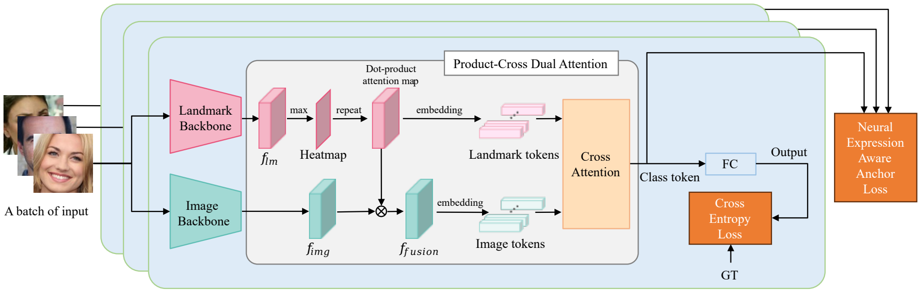


Fig. 1. Overview of our method. Given a face image, our method first extracts landmark features and image appearance features by existing backbone networks. After that, we propose a product-cross dual attention module to fuse the two kinds of features. After the attention module, we obtain a vector of class token, which is finally input to a fully-connected (FC) classifier head to output the expression type of the input face. To optimize the above model, besides the cross-entropy loss, we also propose a neutral expression aware anchor loss applied to the class tokens of all of the samples in a training batch.

Zhang et al. [55] utilized annotator information to calculate triplet loss, introducing a hierarchical structure to construct more refined triplets based on existing triplets, which were used for fine-tuning the network. Cai et al. [3] proposed a island loss function to to extract discriminative features. Li et al. [24] introduced a locality-preserving loss to minimize the distances between samples and their surrounding K samples, preserving the local structure of each sample while maintaining compactness among samples of the same expression. Furthermore, Li et al. [22] proposed an AdaReg loss, which adaptively adjusts expression weights based on the number of different expressions within each batch. This approach addresses class imbalance issues and enhances the discriminative capability of expression representations.

3 Our Method

Figure 1 shows the overview of our method. Given an input face image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, we use a landmark backbone to extract the landmark feature map $f_{lm} \in \mathbb{R}^{h\times w\times c_{lm}}$, where h and w are the height and width of the feature map which has c_{lm} channels, respectively. In this paper, $h = w = 14$, and $c_{lm} = 128$. At the same time, we use an image backbone to extract the image feature map $f_{img} \in \mathbb{R}^{h\times w\times c_{img}}$ ($c_{img} = 256$). The landmark backbone used in this paper is the MobileFaceNet [5], and the image backbone adopted is IR50 [10].

We extract both landmark features and image features because we would like to use the landmark features to guide the learning of the image features, i.e., using the landmark information to enforce the network to focus on the image features around the most prominent landmark regions.

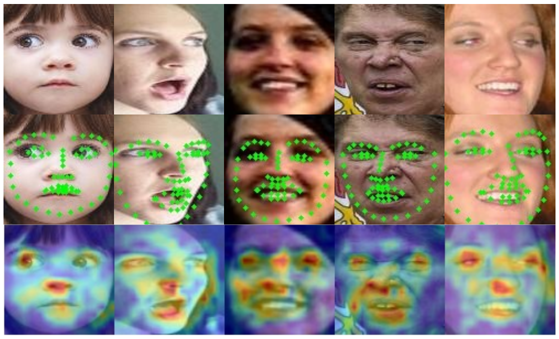


Fig. 2. The visualization of original images (top), images with landmarks (middle) and images with heatmaps (bottom). The facial landmarks are detected by [5]. As we can observe, guided by the heatmap generated for landmark features, the network pays closer attention to crucial facial features such as the eyes, nose, and mouth.

To achieve the above goal, we send both the landmark and image features into the proposed product-cross dual attention module, by which we achieve the fusion of the two kinds of information. After the attention module, we obtain a class token of size c_t (e.g., 768). The class token is then fed into a FC (fully-connected) layer to output the expression class scores of the input image. Each sample in a training batch undergoes the same feature extraction process to output a class token. We compute neutral expression aware anchor loss over all the class tokens of the training batch.

In the following, we elaborate the attention module and the anchor loss (and other losses used to train our model) in detail.

3.1 Product-Cross Dual Attention Module

The proposed product-cross dual attention mechanism in this paper aims at leveraging the positional information provided by facial landmarks. The landmarks can help assign larger weight to important facial regions (such as eyes and mouth) while smaller weight to the unimportant regions (such as hair and background). To achieve this, our product-cross dual attention module uses the landmark features as Query for calculating cross attention. In contrast to traditional self-attention mechanisms, cross attention focuses more on the relationship between the input landmark features and facial features, enabling the model to capture more semantic information about crucial areas. The product-cross dual attention module combines two kinds of attentions: one is the dot-product attention, and the other is the cross-feature attention, which are elaborated in detail as following.

Dot-Product Attention By the dot-product attention, our aim is to enhance the image features around the landmarks. To this end, we first compute a heatmap $\mathbf{H} \in \mathbb{R}^{h \times w \times 1}$ by applying the max pooling operation to the

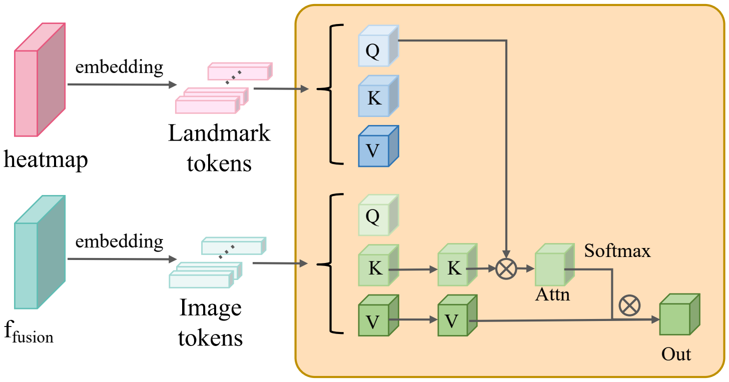


Fig. 3. Illustration of the cross-attention mechanism. We compute both the query, key, and value embeddings for the landmark and fusion tokens. However, we use the queries of landmark, and keys and values of fusion tokens to perform the cross-attention.

landmark feature f_{lm} along the channel dimension. As shown in Figure 2, the values of heatmap around the face landmarks are larger than other places. Than the heatmap is repeated c_{img} times along the channel dimension to obtain $\mathbf{H}' \in \mathbb{R}^{h \times w \times c_{img}}$. Finally, we multiply the heatmap attention map and the image feature map f_{img} in an element-wise manner to obtain the result of the dot-product attention mechanism, i.e.,

$$f_{fusion} = \mathbf{H}' \otimes f_{img}, \tag{1}$$

where $f_{fusion} \in \mathbb{R}^{h \times w \times c_{img}}$, and $\otimes$ represents the element-wise multiplication. After the dot-product attention, the image features around the landmarks are enhanced (such as eye and mouth regions), while the features at other places are weakened (such as hair regions and background).

Cross-Feature Attention After obtaining the landmark feature and fusion feature, we proceed to embed them into landmark tokens $t_{lm} \in \mathbb{R}^{(hw) \times c_t}$ and fusion tokens $t_{fusion} \in \mathbb{R}^{(hw) \times c_t}$. Additionally, we introduce a learnable class token in $\mathbb{R}^{1 \times c_t}$ to represent global features (this class token is ultimately fed into a fully connected layer for expression classification), resulting in full landmark tokens $t_{lm} \in \mathbb{R}^{(hw+1) \times c_t}$ and fusion tokens $t_{fusion} \in \mathbb{R}^{(hw+1) \times c_t}$.

$$\begin{aligned} t_{lm} &= Cat(Embedding(f_{lm}), t_{lm\_class}), \\ t_{fuison} &= Cat(Embedding(f_{fuison}), t_{fusion\_class}), \end{aligned} \tag{2}$$

where t_{lm_class} and $t_{fuison\_class} \in \mathbb{R}^{1 \times c_t}$ are landmark class token and fusion class token, respectively. t_{lm} and t_{fuison} are landmark tokens and fusion tokens, respectively, which are inputted into the subsequent transformer blocks.

We employ a transformer to compute relationships between tokens. The process of cross-feature attention is illustrated in Figure 3. To begin with, the landmark features and fusion features are mapped into three matrices each: a fusion query matrix Q_{fusion}, a fusion key matrix K_{fusion}, and a fusion value matrix V_{fusion}, as well as a landmark query matrix Q_{lm}. The expressions for this mapping are as follows:

$$\begin{aligned} Q_{fusion} &= t_{fusion} \times W_{q\_fusion}, \\ K_{fusion} &= t_{fusion} \times W_{k\_fusion}, \\ V_{fusion} &= t_{fusion} \times W_{v\_fusion}, \\ Q_{lm} &= t_{lm} \times W_{q\_lm}, \end{aligned} \tag{3}$$

where W_{q_fusion}, W_{k_fusion}, W_{v_fusion}, $W_{q\_lm} \in \mathbb{R}^{c_t \times c_t}$ are the mapping matrices.

Then, we calculate the cross-attention between the landmark query matrix Q_{lm} and the fusion key matrix K_{fusion}, along with the fusion value matrix V_{fusion}. This process can be mathematically described as follows:

$$CrossAttention(Q_{lm}, K_{fusion}, V_{fusion}) = Softmax(\frac{Q_{lm}K_{fusion}^T}{\sqrt{d}})V_{fusion}, \tag{4}$$

where $Softmax(\cdot)$ is softmax activation function and $\sqrt{d}$ is the scaling factor for normalization.

Using the landmark query matrix Q_{lm} instead of the fusion query matrix Q_{fusion}, is done to make better use of the spatial positional information contained within the landmark feature. This helps guide and focus on the regions within the fusion feature that are more relevant to the expression being conveyed. Subsequently, we calculate the output of a transformer block t_{fusion_out}. The t_{fusion_out} is of the same size as t_{fusion}.

$$\begin{aligned} t'_{fusion} =& CrossAttention(Q_{lm}, K_{fusion}, V_{fusion}) + t_{fusion}, \\ t_{fusion\_out} =& MLP(Norm(t'_{fusion}) + t'_{fusion}, \end{aligned} \tag{5}$$

where $MLP(\cdot)$ is multi-layer perceptron, $Norm(\cdot)$ is normalization function.

After passing through several transformer blocks, the class token is putted into the FC to calculate the final classification prediction.

3.2 Neutral Expression Aware Anchor Loss

Due to the inherent challenge of facial expressions characterized by intra-class variation and inter-class similarity, several loss functions, such as center loss, triplet loss, anchor loss, and others, have been proposed to minimize the feature distances among samples of the same class while increasing distances between samples of different classes. These loss functions treat all expressions uniformly. However, in practice, neutral expressions differ from other expressions.

In a continuous expression representation space like arousal-valence model (as shown in Figure 4), neutral expressions exhibit no distinction on arousal or valence, meaning that the extracted features for neutral expressions should

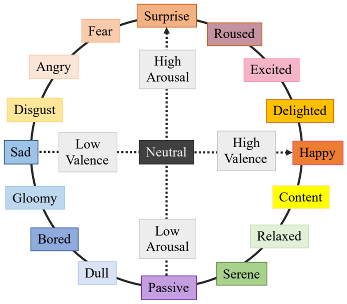


Fig. 4. The Pleasure Arousal Dominance Emotion Model [31]: Discrete emotions mapped into a 2D coordinate space of arousal and valence [12]. Mittal et al. [31] mapped the emotion labels to discrete emotions in this 2D space.

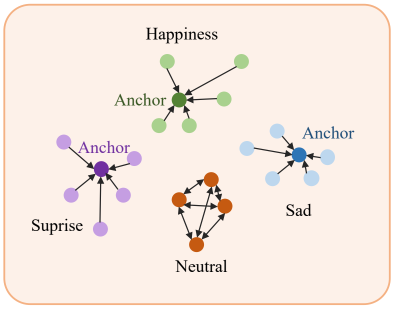


Fig. 5. Illustration of neutral-expression-aware anchor loss. For expressions other than "neutral", we find an anchor and reduce the distance between the training samples to the anchor. For neutral expression, we directly minimize the distance between any pair of neutral samples, as we argue that all the neutral expressions own the same degree of expressiveness without arousal-valence variance.

be very close in the feature space, with minimal variance. In contrast, other expressions such as happy, sad, etc., exhibit different degrees, i.e., there are very happy or just a little happy expressions.

The above insight inspires that we need to handle neutral and non-neutral expressions differently. Therefore, we propose an improvement to the anchor loss, introducing the Neutral-Expression-Aware Anchor Loss (see Figure 5).

Firstly, let us define the traditional anchor loss. For a batch of samples, the anchor loss function first identifies an anchor sample for each class of expressions:

$$anchor\_c = \underset{i \in N_c}{\arg\min}\, \text{Confidence}(f_i), \tag{6}$$

where $anchor\_c$ represents the index of the anchor sample of the c^{th} expression class, $i \in N_c$ indexes all the samples in the training batch with expression class c, and Confidence$(\cdot)$ is the formula used for calculating sample confidence which is computed as the entropy of the predicted expression classification scores by the FC classification layer. f_i is the final feature of the sample output by the attention module, i.e., the class token. In total, Eq. 6 finds for each class the most confident sample and returns the index of the sample.

With the anchors for different expression class, the anchor loss function calculates the loss as the distance between other samples of the same class and the anchor sample, as shown in the following formula:

$$L_{anchor} = \frac{1}{N_c} \sum_{c=0}^{C} \sum_{i=0}^{N_c - 1} \text{Dist}(f_i, f_{anchor\_c}) \tag{7}$$

where C is the number of all the expression classes, N_c represents the number of samples from the c-th expression class in the batch, f_i is i-th sample in c-th class, f_{anchor_c} is the anchor sample in this class. Dist($\cdot$) is the formula used for distance calculation, and in this paper, we employ the mean square error function.

Now we define our proposed neutral-aware anchor loss. It composes of two parts. The first part computes the above anchor loss but excludes the neutral expression class.

$$L_{non-neutral} = \frac{1}{N_c} \sum_{c=0}^{C'} \sum_{i=0}^{N_c-1} \mathrm{Dist}(f_i, f_{anchor\_c}) \tag{8}$$

where C' is the set of all the expression classes except the neutral class. As for neutral expressions in the training batch, we impose a stricter constraint by requiring all sample features to be equal. Through this more rigorous constraint, we aim to make their distribution in the feature space converge towards a single point. The complete description of the loss of the neutral class is as follows:

$$L_{neutral} = \frac{1}{(N_n - 1)^2} \sum_{i=0}^{N_n-1} \sum_{j=0, j\neq i}^{N_n-1} \mathrm{Dist}(f_i, f_j), \tag{9}$$

where N_n represents the number of samples from the neutral expression class in the batch.

Finally, the Neutral-Aware Anchor (NeAA) loss is defined as:

$$L_{NeAA} = L_{non-neutral} + L_{neutral}. \tag{10}$$

3.3 Total Loss Function

In the proposed model, the image backbone, the landmark backbone, and the cross-attention module are jointly trained in an end-to-end fashion. We calculate the cross-entropy loss L_{cls} for the final classification results. Overall, the total loss in the training of the entire network is as follows:

$$L = L_{cls} + \lambda L_{NeAA}, \tag{11}$$

where the hyper-parameter $\lambda = 0.01$ is used to balance the loss function.

4 Experiments

4.1 Datasets

RAF-DB: The Real-world Affective Face Database (RAF-DB) [24] is a large-scale database, which includes 29,672 real-world facial images collected by searching on Flickr. The images are with great variability in age, ethnicity, lighting conditions, etc. With manually crowd-sourced annotation and reliable estimation, RAF-DB provides 7 basic expression classes (happiness, surprise, sadness, anger, disgust, fear, and neutral). For facial expression recognition task, there are 15,339 facial expression images utilized (12,271 images are used for training and 3,068 images are used for testing).

FERPlus: The FERPlus dataset [1] is extended from FER2013 [16] used in the ICML 2013 Challenge. FER2013 is a large-scale dataset collected by APIs in the Google search, which includes images resized to 48X48 pixels. It contains 28709 training images, 3589 validation images and 3589 test images. It is relabeled in 2016 by Microsoft with each image labeled by 10 individuals to consist 8 classes (7 basic expressions and contempt expression), thus has more reliable annotations.

AffectNet: AffectNet [32] is one of the largest datasets in the wild, containing over a million images collected from the Internet by querying various search engines. It provides two facial expression models (categorical model and dimensional model). For the FER task, there are a total of 420K images manually annotated into eight classes of expressions. Following the setup in [54], we used 280K training images and 3500 validation images (500 images per category) with 7 expression categories.

4.2 Implementation Details

All images are resized to 112×112 pixels before feeding into the model. We initialize the image backbone with IR50 [10], pretrained on the Ms-Celeb-1M dataset [17], and use MobileFaceNet [5] as the landmark backbone. During training, we keep the landmark backbone parameters fixed to ensure the accuracy of landmark information. For the cross-attention module, we employ a ViT with 2 transformer blocks. The MLP ratio is set to 4, and the drop ratio is 0.5. Our model is trained for 200 epochs using the Adam optimizer, with a batch size of 144. For the RAF-DB and FERPlus datasets, the learning rate is set to 3.5e-5, while for the AffectNet dataset, it is set to 1e-6. To augment the training data, we apply random horizontal flips and random erasing, while for the testing data, we only perform resize operation. Our model is implemented using PyTorch [35] and trained on two NVIDIA RTX 3090 GPUs.

4.3 Ablation Study

We conducted ablation study on the RAF-DB dataset to investigate the impact of the model's architecture, various proposed modules, and loss functions.

Effectiveness of Dot-Product Attention and Cross-feature Attention. To evaluate the impact of the modules proposed in this paper, we conducted an ablation study on the RAF-DB dataset to investigate the effects of Dot-Product Attention and Cross-feature Attention on the final classification results. As shown in Table 1, it is evident that the addition of these modules leads to an overall improvement in accuracy on the validation set. After incorporating the Dot-Product Attention module into the model, the accuracy increased from 91.67% (row 6) to 92.31% (row 8), marking a 0.64% improvement. Similarly, with the inclusion of the Cross-feature Attention module, the accuracy on the test set also increased by 0.36% (from row 7 to row 8). With the use of both

Table 1. Evaluation (%) of Dot-Product Attention, Cross-Feature Attention and NeAA Loss on RAF-DB.

	Dot_attn	Cross_attn	NeAA Loss	RAF-DB
1				91.30
2	✓			91.42
3		✓		91.49
4			✓	91.65
5	✓	✓		91.79
6		✓	✓	91.67
7	✓		✓	91.95
8	✓	✓	✓	**92.31**

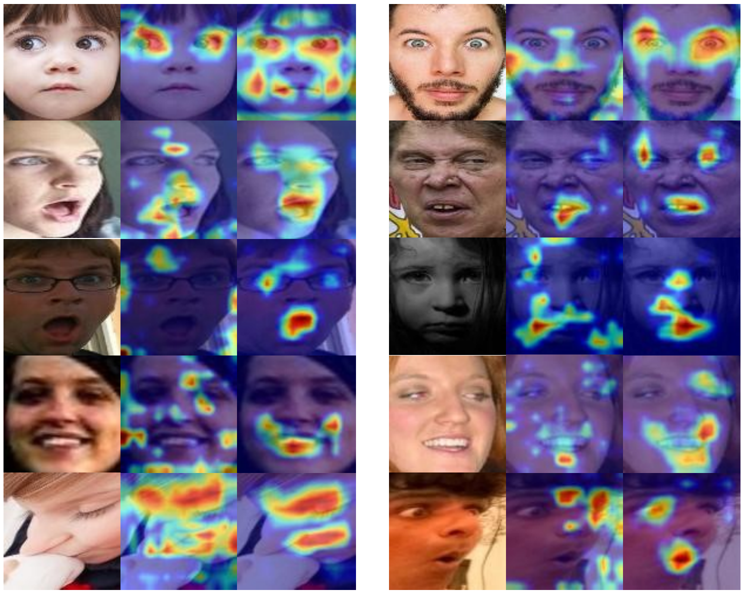


Fig. 6. The visualization of original images (left column), features without (middle column) and with Product-Cross Dual Attention Module (right column). The results show that the model emphasizes regions that significantly represent facial expressions.

Attention modules, the accuracy increased from 91.65% (row 4) to 92.31% (row 8). We employed Grad-CAM to visualize the features after the application of the Product-Cross Dual Attention Module. Figure 6 shows that the model emphasizes regions that significantly represent facial expressions, such as the mouth in surprised expressions (row 2 and row 3 on the left half) and happy expressions (row 4), and the eyes in surprised expressions (row 1 and row 5 on the right half). Furthermore, due to the implicit inclusion of head pose information in landmark features, the model can better extend to facial expression recognition

Table 2. Evaluation (%) of pre-trained ViT model on RAF-DB.

Pre-trained ViT	#Param	#FLOPs	RAF-DB
✓	67.3M	17.9G	91.88
	34.2M	9.3G	**92.31**

in real-world scenarios with various head poses. As illustrated in the last row of Figure 6, even under substantial head pose variations, the model remains capable of accurately identifying the positions of important areas in facial images.

We also experimentally explore whether the pre-trained large ViT model can help improve the accuracy of expression recognition. The Cross-feature Attention network is initialized with the parameters of ViT pre-trained on ImageNet [9]. As shown in Table 2, the pre-trained ViT model did not yield improvement in accuracy. The possible reason is that existing expression recognition datasets are small, and deeper networks increase the risk of overfitting. Therefore, a simplified ViT model with only 2 blocks is deemed sufficient. This not only ensures recognition accuracy but also reduces the consumption of computing resources and time.

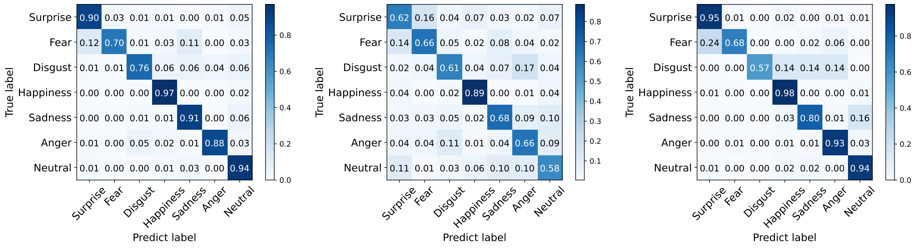
Effectiveness of Neutral Expression Aware Loss. To validate the impact of the proposed Neutral Expression Aware Loss (NeAA Loss), we compared the accuracy of models trained with and without NeAA Loss on the RAF-DB test set. As shown in Table 1, the first row and the fourth row indicate that NeAA Loss is beneficial for improving model accuracy. The confusion matrix in Figure 7 (a) also reveals that only a few instances of other expressions are misclassified as neutral expressions. The Recall and F1-score for neutral expressions with and without the use of NeAA Loss is shown in Table 3. It is evident that the model's ability to recognize neutral expressions improves when the NeAA Loss is applied. Additionally, it enhances the discriminative ability between neutral expressions and other expressions with low arousal level. As shown in Figure 8, many neutral expressions and other expressions with low arousal are incorrectly predicted (red labels) when there is no NeAA loss, while with the application of NeAA Loss, these predictions are corrected (black labels).

Table 3. Recall and F1-score for neutral expressions with and without the use of NeAA Loss on RAF-DB.

NeAA Loss	Recall	F1-score
	0.8971	0.9050
✓	**0.9397**	**0.9116**

Table 4. Evaluation (%) of different methods that generate the dot-product attention that is multiplied with the image feature f_{img}, on RAF-DB.

	Method	RAF-DB
1	Conv	91.46
2	Sum+Repeat	91.72
3	Abs+Sum+Repeat	91.59
4	Max+Repeat	**92.31**



(a) (b) (c)

Fig. 7. Confusion matrices of our model on RAF-DB (subfigure (a)), AffectNet(7cls) (subfigure (b)) and FERPlus datasets (subfigure (c)). Our method exhibits clear and strong performance in terms of class-wise accuracy (diagonals of each confusion matrix) across all three datasets.

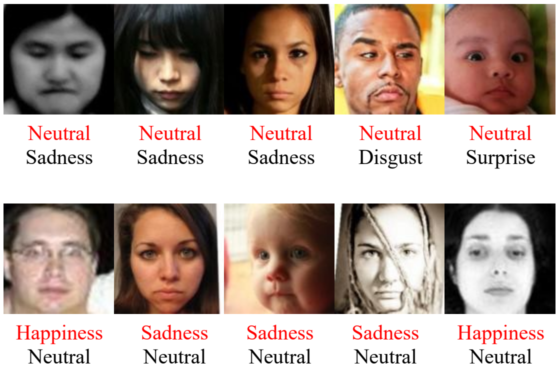


Fig. 8. Misidentified expressions in low arousal levels (top) and neutral expressions (bottom). The wrong predictions without NeAA are highlighted in red. With the application of the NeAA Loss, these predictions are corrected (black labels).

Effectiveness of Different Dot-Product Attention Map Generation Methods. There are various methods for generating the dot-product attention map (see Figure 1) that is multiplied with the image feature map. We conducted an ablation study on these methods using the RAF-DB dataset to assess their impact on the final results. As shown in Table 4, we generated the attention map from landmark features using Conv (a convolutional layer that directly maps the landmark feature f_{lm} in space $\mathbb{R}^{14\times14\times128}$ to the dot-product attention map in space $\mathbb{R}^{14\times14\times256}$), Sum+Repeat (sum the f_{lm} along the channel dimension to obtain a feature map in space $\mathbb{R}^{14\times14\times1}$ and then repeat it 256 times along the channel dimension to obtain the attention map), Abs+Sum+Repeat (compute abs(f_{lm}) at first, then sum it along the channel dimension, and finally repeat 256 times), and Max+Repeat (compute max pooling of f_{lm} along the channel dimension and then repeat) methods. Among these, the Max+Repeat method achieved the best result (92.31%). The convolution method performed worse

Table 5. Performance comparison (%) with SOTA methods on RAF-DB, AffectNet(7cls) and FERPlus datasets.

Method	Year	RAF-DB	AffectNet(7cls)	FERPlus
SCN [46]	CVPR 2020	87.03	-	89.39
PSR [45]	CVPR 2020	88.98	63.77	-
RAN [47]	TIP 2020	86.90	-	89.16
DACL [14]	WACV 2021	87.78	65.20	-
KTN [22]	TIP 2021	88.07	63.97	90.49
DMUE [42]	CVPR 2021	89.42	63.11	-
FDRL [39]	CVPR 2021	89.47	-	-
ARM [43]	arXiv 2021	90.42	65.20	-
TransFER [50]	ICCV 2021	90.91	66.23	90.83
APViT [51]	CVPR 2022	91.98	66.91	90.86
Meta-Face2Exp [52]	CVPR 2022	88.54	64.23	-
EAC [56]	ECCV 2022	89.99	65.32	89.64
RANet [29]	FG 2023	89.57	65.09	-
SwinFace [38]	TCSVT 2023	90.97	-	-
Latent-OFER [21]	ICCV 2023	89.6	63.9	-
POSTER [57]	ICCV 2023	92.05	**67.31**	91.62
Ours	-	**92.31**	67.14	**92.97**

Table 6. Per-class performance comparison (%) with POSTER on RAF-DB and AffectNet(7cls) datasets.

Dataset	Method	Neutral	Happy	Sad	Surprise	Fear	Disgust	Anger	mean Acc
RAF-DB	POSTER	92.35	**96.96**	**91.21**	**90.27**	67.57	75.00	**88.89**	86.04
RAF-DB	Ours	**93.97**	**96.96**	91.00	89.67	**70.27**	**75.62**	87.65	**86.45**
AffectNet(7cls)	POSTER	**67.20**	**89.00**	67.00	**64.00**	64.80	56.00	62.60	**67.23**
AffectNet(7cls)	Ours	58.20	88.60	**68.40**	62.40	**66.00**	**60.80**	**65.60**	67.14

than all the other operators, possibly due to significant alterations in the original landmark information caused by convolution operations, leading to a change in the network's focus area.

4.4 Comparison with the State-of-the-Art Methods

We compared the proposed method in this paper with some state-of-the-art (SOTA) methods on the RAF-DB, FERPlus, and AffectNet datasets, and the results are presented in Table 5.

Results on RAF-DB. The results of the comparison with the SOTA methods on the RAF-DB dataset are shown in Table 5, in the 3-th column. Our proposed method outperforms all the compared methods in terms of accuracy (accuracy across all samples), achieving an accuracy of 92.31%, which is 0.26% higher than the second-best method, POSTER. We conducted an analysis in Table 6 comparing the accuracies of our method and POSTER on each class in the RAF-DB dataset. It is evident that our method achieved a higher accuracy on the neutral, fear, disgust expression compared to POSTER, but had a relatively lower accuracy on the Anger expression.

Results on AffectNet. Since the test set of the AffectNet dataset is not publicly available, we conducted our comparison on the validation set following SOTA methods. Due to the extreme class imbalance in the AffectNet data, we applied oversampling techniques similar to RAN, POSTER, and APViT. We compared the accuracy of different methods on the 7-class emotion recognition task in the AffectNet dataset. From the 4-th column of Table 5, it can be seen that our method achieved an accuracy of 67.14%. While not the highest, it secured the second position. Our result is 0.17% lower than POSTER's results, probably because we do not process the feature in a multi-resolution manner, while POSTER performs that by employing a pyramid network structure. However, in terms of running time per image, POSTER takes 3ms (see Table 7), while our model only takes 1.3ms. Our model achieves more efficient expression recognition with a small loss of accuracy. We conducted a detailed analysis comparing our method and POSTER's accuracy on each emotion class in Table 6. Our method achieved the best results on the sad, fear, disgust and anger emotions, outperforming POSTER a lot. However, the recognition accuracy on the neutral emotion was relatively lower. It is also worth noting that the confusion matrix for AffectNet (see Figure 7 (b)) indicates an improved ability of our model to differentiate between neutral and other emotions.

Results on FERPlus. The results of the comparison with the SOTA methods on the FERPlus dataset are displayed in Table 5, in the 5-th column. Our method achieved an accuracy of 92.97%, surpassing the second-best method, POSTER, by 1.35%. The confusion matrix for FERPlus is illustrated in Figure 7 (c), which reveals that we have less error in neutral expression and other expressions.

4.5 Comparison on Number of Parameters and Running Performance

In Table 7, we compare our method with the SOTA approaches on the number of parameters and the FLOPs. As can be seen, our method not only outperforms other methods in terms of recognition accuracy, but also uses less network parameters and runs faster than the SOTA approaches. Compared to models with similar structures such as APViT and POSTER, our model incorporates MobileFaceNet, a lightweight and efficient landmark detector, into the network. Moreover, we reduce the Cross-Attention Module to only 2 blocks, aiming to simplify the model while maintaining effectiveness. This design choice facilitates better scalability in real-time or resource-constrained environments. Table 8 reveals the number of parameters and FLOPs for each module in our model, showing that the added Landmark Backbone has minimal impact on model complexity and the number of parameters. We conducted test on a single GPU, and the average running time taken for each method per image is shown in Table 7. Our model

Table 7. Comparison on Parameter Number and FLOPs. The image backbone (IR50) and facial landmark detector (MobileFaceNet) are taken into account when computing Params and FLOPs.

Methods	#Param	#FLOPs	RAF-DB	AffectNet	Running time
DMUE [42]	78.4M	13.4G	89.42	63.11	-
TransFER [50]	65.2M	15.3G	90.91	66.23	-
APViT [51]	-	12.7G	91.98	66.91	3.9ms
POSTER [57]	71.8M	15.7G	92.05	**67.31**	3.0ms
Ours	**34.2M**	**9.3G**	**92.31**	67.14	**1.3ms**

Table 8. Parameter Number and FLOPs of each module in our method.

Module	#Param	#FLOPs
Image Backbone	17.6M	5.5G
Landmark Backbone	1.0M	0.2G
Cross Attention Module	15.6M	3.6G

demonstrates shorter inference time compared to APViT and POSTER, being only 1/3 of APViT and 2/5 of POSTER. This indicates that we can use more concise network architecture to fulfill the FER task if more effective data processing modules are designed and adopted such as the proposed product-cross dual attention method and the neutral-expression-aware anchor loss function.

5 Conclusion

This paper proposes a simple yet effective face expression recognition method. The two main contributions of this paper are that we propose a product-cross dual attention mechanism and a neutral-expression-aware anchor loss. With the dual attention module, we combine the features extracted by the landmark backbone and image backbone. The features around the position of landmarks are successfully enhanced, while reducing the influence of features at other places. This indicates that making the network focus on important regions is useful and can indeed improve recognition accuracy. The neutral-aware loss takes the special characteristic of neutral expressions into consideration, i.e., all the neutral features should be similar to each other. With this constraint, we further improve the recognition accuracy by constraining the learning space of the network meaningfully. We have conducted comparison and ablation experiments which validate the effectiveness of our method. In particular, our method outperforms the latest method POSTER while running faster.

Acknowledgements. The work is supported by Guangdong International Technology Coopertation Project (No.2022A0505050009), and National Science Foundation of China (No. 62072191, 61972160).

References

1. Barsoum, E., Zhang, C., Ferrer, C.C., Zhang, Z.: Training deep networks for facial expression recognition with crowd-sourced label distribution. In: Proceedings of the 18th ACM international conference on multimodal interaction. pp. 279–283 (2016)
2. Bashyal, S., Venayagamoorthy, G.K.: Recognition of facial expressions using gabor wavelets and learning vector quantization. Engineering Applications of Artificial Intelligence **21**(7), 1056–1064 (2008)
3. Cai, J., Meng, Z., Khan, A.S., Li, Z., O'Reilly, J., Tong, Y.: Island loss for learning discriminative features in facial expression recognition. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). pp. 302–309. IEEE (2018)
4. Chang, Y., Wang, S.: Knowledge-driven self-supervised representation learning for facial action unit recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20417–20426 (2022)
5. Chen, C.: PyTorch Face Landmark: A fast and accurate facial landmark detector (2021), https://github.com/cunjian/pytorch_face_landmark
6. Chen, J., Gao, C., Sun, L., Sang, N.: Ccsd: cross-camera self-distillation for unsupervised person re-identification. Visual Intelligence **1**(1), 27 (2023)
7. Cheng, R., Wang, X., Sohel, F., Lei, H.: Topology-aware universal adversarial attack on 3d object tracking. Visual Intelligence **1**(1), 1–12 (2023)
8. Darwin, C., Prodger, P.: The expression of the emotions in man and animals. Oxford University Press, USA (1998)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
10. Deng, J., Guo, J., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition (2018)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
12. Ekman, P., Friesen, W.V.: Head and body cues in the judgment of emotion: A reformulation. Perceptual and motor skills **24**(3), 711–724 (1967)
13. Ekman, P., Friesen, W.V.: Constants across cultures in the face and emotion. Journal of personality and social psychology **17**(2), 124 (1971)
14. Farzaneh, A.H., Qi, X.: Facial expression recognition in the wild via deep attentive center loss. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2402–2411 (2021)
15. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., et al.: Domain-adversarial training of neural networks. The journal of machine learning research **17**(1), 2096–2030 (2016)
16. Goodfellow, I.J., Erhan, D., Carrier, P.L., Courville, A., Mirza, M., Hamner, B., et al.: Challenges in representation learning: A report on three machine learning contests. In: Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20. pp. 117–124. Springer (2013)
17. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14. pp. 87–102. Springer (2016)

18. Jung, H., Lee, S., Yim, J., Park, S., Kim, J.: Joint fine-tuning in deep neural networks for facial expression recognition. In: Proceedings of the IEEE international conference on computer vision. pp. 2983–2991 (2015)
19. Kollias, D.: Multi-label compound expression recognition: C-expr database & network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5589–5598 (2023)
20. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
21. Lee, I., Lee, E., Yoo, S.B.: Latent-ofer: Detect, mask, and reconstruct with latent vectors for occluded facial expression recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1536–1546 (2023)
22. Li, H., Wang, N., Ding, X., Yang, X., Gao, X.: Adaptively learning facial expression representation via cf labels and distillation. IEEE Transactions on Image Processing **30**, 2016–2028 (2021)
23. Li, P., Sun, H., Huang, C., Shen, J., Nie, Y.: Interactive image/video retexturing using gpu parallelism. Computers & Graphics **36**(8), 1048–1059 (2012)
24. Li, S., Deng, W., Du, J.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2852–2861 (2017)
25. Li, X., Deng, W., Li, S., Li, Y.: Compound expression recognition in-the-wild with au-assisted meta multi-task learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5734–5743 (2023)
26. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proceedings of the seventh IEEE international conference on computer vision. vol. 2, pp. 1150–1157. Ieee (1999)
27. Ma, F., Sun, B., Li, S.: Facial expression recognition with visual transformers and attentional selective fusion. IEEE Transactions on Affective Computing (2021)
28. Ma, T., Nie, Y., Zhang, Q., Zhang, Z., Sun, H., Li, G.: Effective video stabilization via joint trajectory smoothing and frame warping. IEEE Transactions on Visualization and Computer Graphics **26**(11), 3163–3176 (2019)
29. Ma, X., Ma, Y.: Relation-aware network for facial expression recognition. In: 2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG). pp. 1–7. IEEE (2023)
30. Meng, Z., Liu, P., Cai, J., Han, S., Tong, Y.: Identity-aware convolutional neural network for facial expression recognition. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). pp. 558–565. IEEE (2017)
31. Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., Manocha, D.: M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 1359–1367 (2020)
32. Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing **10**(1), 18–31 (2017)
33. Nie, Y., Zhang, Q., Wang, R., Xiao, C.: Video retargeting combining warping and summarizing optimization. The Visual Computer **29**, 785–794 (2013)
34. Pan, Y., Niu, Z., Wu, J., Zhang, J.: Insocialnet: Interactive visual analytics for role—event videos. Computational Visual Media **5**, 375–390 (2019)
35. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., et al.: Automatic differentiation in pytorch (2017)

36. Peng, Z., Jiang, B., Xu, H., Feng, W., Zhang, J.: Facial optical flow estimation via neural non-rigid registration. Computational Visual Media **9**(1), 109–122 (2023)
37. Pons, G., Masip, D.: Multi-task, multi-label and multi-domain learning with residual convolutional networks for emotion recognition. arXiv preprint arXiv:1802.06664 (2018)
38. Qin, L., Wang, M., Deng, C., Wang, K., Chen, X., Hu, J., et al.: Swinface: A multi-task transformer for face recognition, expression recognition, age estimation and attribute estimation. IEEE Transactions on Circuits and Systems for Video Technology (2023)
39. Ruan, D., Yan, Y., Lai, S., Chai, Z., Shen, C., Wang, H.: Feature decomposition and reconstruction learning for effective facial expression recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7660–7669 (2021)
40. Ryou, S., Jeong, S.G., Perona, P.: Anchor loss: Modulating loss scale based on prediction difficulty. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5992–6001 (2019)
41. Shan, C., Gong, S., McOwan, P.W.: Facial expression recognition based on local binary patterns: A comprehensive study. Image and vision Computing **27**(6), 803–816 (2009)
42. She, J., Hu, Y., Shi, H., Wang, J., Shen, Q., Mei, T.: Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6248–6257 (2021)
43. Shi, J., Zhu, S., Liang, Z.: Learning to amend facial expression representation via de-albino and affinity. arxiv 2021. arXiv preprint arXiv:2103.10189
44. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., et al.: Attention is all you need. Advances in neural information processing systems **30** (2017)
45. Vo, T.H., Lee, G.S., Yang, H.J., Kim, S.H.: Pyramid with super resolution for in-the-wild facial expression recognition. IEEE Access **8**, 131988–132001 (2020)
46. Wang, K., Peng, X., Yang, J., Lu, S., Qiao, Y.: Suppressing uncertainties for large-scale facial expression recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6897–6906 (2020)
47. Wang, K., Peng, X., Yang, J., Meng, D., Qiao, Y.: Region attention networks for pose and occlusion robust facial expression recognition. IEEE Transactions on Image Processing **29**, 4057–4069 (2020)
48. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14. pp. 499–515. Springer (2016)
49. Xiao, C., Nie, Y., Hua, W., Zheng, W.: Fast multi-scale joint bilateral texture upsampling. The Visual Computer **26**, 263–275 (2010)
50. Xue, F., Wang, Q., Guo, G.: Transfer: Learning relation-aware facial expression representations with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3601–3610 (2021)
51. Xue, F., Wang, Q., Tan, Z., Ma, Z., Guo, G.: Vision transformer with attentive pooling for robust facial expression recognition. IEEE Transactions on Affective Computing (2022)
52. Zeng, D., Lin, Z., Yan, X., Liu, Y., Wang, F., Tang, B.: Face2exp: Combating data biases for facial expression recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20291–20300 (2022)

53. Zhang, K., Huang, Y., Du, Y., Wang, L.: Facial expression recognition based on deep evolutional spatial-temporal networks. IEEE Transactions on Image Processing **26**(9), 4193–4203 (2017)
54. Zhang, Q., Nie, Y., Zhu, L., Xiao, C., Zheng, W.S.: A blind color separation model for faithful palette-based image recoloring. IEEE Transactions on Multimedia **24**, 1545–1557 (2021)
55. Zhang, W., Ji, X., Chen, K., Ding, Y., Fan, C.: Learning a facial expression embedding disentangled from identity. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6759–6768 (2021)
56. Zhang, Y., Wang, C., Ling, X., Deng, W.: Learn from all: Erasing attention consistency for noisy label facial expression recognition. In: European Conference on Computer Vision. pp. 418–434. Springer (2022)
57. Zheng, C., Mendieta, M., Chen, C.: Poster: A pyramid cross-fusion transformer network for facial expression recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3146–3155 (2023)
58. Zhi, R., Flierl, M., Ruan, Q., Kleijn, W.B.: Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) **41**(1), 38–52 (2010)
59. Zhong, L., Liu, Q., Yang, P., Liu, B., Huang, J., Metaxas, D.N.: Learning active facial patches for expression analysis. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 2562–2569. IEEE (2012)