Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Constrained and directional ensemble attention for facial action unit detection

Zhiwen Shao [a,b,c,d], Bikuan Chen [a,b,*], Yong Zhou [a,b], Xuehuai Shi [e], Canlin Li [f], Lizhuang Ma [d], Dit-Yan Yeung [c]

[a] School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China
[b] Mine Digitization Engineering Research Center of the Ministry of Education, Xuzhou 221116, China
[c] Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon 999077, Hong Kong, China
[d] School of Computer Science, Shanghai Jiao Tong University, Shanghai 200240, China
[e] School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
[f] School of Computer Science and Technology, Zhengzhou University of Light Industry, Zhengzhou 450002, China

ARTICLE INFO

Dataset link: http://www.cs.binghamton.edu/~lijun/Research/3DFE/3DFE_Analysis.html, http://mohammadmahoor.com/pages/databases/disfa, https://osf.io/7wcyz

Keywords:
Facial action unit detection
Self-attention constraining
Dual-directional attention
Multi-view ensemble

ABSTRACT

Facial action unit (AU) detection is a challenging task, due to the subtlety of each AU in local area and the correlations among AUs in global face. In recent years, the prevailing attention mechanism has been introduced to AU detection. However, the inherent mechanism of self-attention weight distribution has been rarely explored. Besides, ensemble learning is an efficient technique, but gains little attention in AU detection. Considering the above limitations, we propose a local self-attention constraining (LSC) network, by regarding the self-attention distribution of each AU as a spatial distribution, and constraining it based on prior knowledge so as to capture AU-related local information. Moreover, to learn correlations among different AU regions, we propose a global dual-directional attention (GDA) network, which adaptively learns global attention map from both vertical and horizontal directions. Last but not least, the two networks from different views of capturing patterns are assembled to integrate both advantages. Extensive experiments on BP4D, DISFA, and GFT benchmarks demonstrate that our methods including local self-attention constraining, global dual-directional attention, and multi-view ensemble all significantly surpass state-of-the-art AU detection works.

1. Introduction

Facial expression is a primary means of conveying human emotions. The analysis and recognition of expressions hold wide-ranging potential in emotion recognition [1] and virtual reality [2]. Defined by the facial action coding system (FACS) [3], a facial action unit (AU) represents local facial movements and describes subtle motions in expressions. AU detection aims to determine the activation status of each AU in a facial image. The identification of AUs remains challenging due to their subtlety and the complexity of their interrelationships.

Human faces have complex structures and can present a wide range of subtle movements, which allow different groups of AUs to convey diverse emotions and intentions. Fig. 1(a) illustrates that AUs with opposite semantics, like AU 12 (lip corner puller) and AU 15 (lip corner depressor), still can co-occur. During the dynamic transition from a smile to a sad expression, an intermediate stage may incorporate such types of AUs. The subtlety of AUs presents difficulties for their identification, necessitating the development of detailed and dynamic modeling techniques. On the other hand, most AUs do not exist independently, and they exhibit complex interrelations and constraints. Certain AUs often occur together. For instance, as shown in Fig. 1(c), a smile typically activates both AU 12 and AU 6 (cheek raiser) simultaneously. The modeling of relationships among AUs is also crucial.

Recently, vision transformers have made significant strides in various computer vision tasks. To accurately detect highly subtle AUs, a few studies have integrated transformers into AU detection to capture the AU relationships [4]. Besides, Li et al.'s work [5] explores the convolutional attention weight distribution, which greatly improves the performance. However, the research about inherent mechanism of transformer attention (also known as self-attention) weight distribution

[*] Corresponding author at: School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China.
E-mail addresses: zhiwen_shao@cumt.edu.cn (Z. Shao), bikuan_chen@cumt.edu.cn (B. Chen), yzhou@cumt.edu.cn (Y. Zhou), xuehuai@njupt.edu.cn (X. Shi), li-cl@zzuli.edu.cn (C. Li), ma-lz@cs.sjtu.edu.cn (L. Ma), dyyeung@cse.ust.hk (D.-Y. Yeung).

https://doi.org/10.1016/j.patcog.2025.111904
Received 20 August 2024; Received in revised form 12 February 2025; Accepted 22 May 2025
Available online 7 June 2025

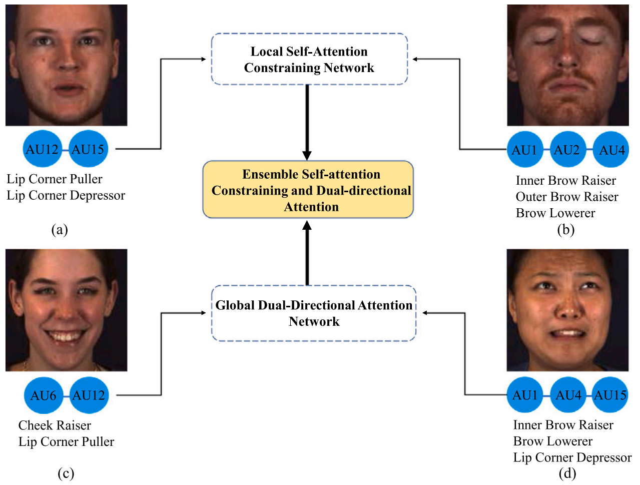


Fig. 1. Example expressions composed of different facial AUs: (a) subtle mouth corner movements via AU 12 (lip corner puller) and AU 15 (lip corner depressor); (b) multiple eyebrow movements depicted by AU 1 (inner brow raiser), AU 2 (outer brow raiser), and AU 4 (brow lowerer); (c) a typical Duchenne smile by combining AU 6 (cheek raiser) with AU 12; (d) a sad expression, conveyed through AU 1, AU 4, and AU 15. The local self-attention constraining (LSC) network captures the subtlety of local AUs, and the globa dual-directional attention (GDA) network focuses on capturing the AU relationships. At last, two networks are assembled as Ensemble Self-attention Constraining and Dual-directional Attention (ESCDA).

is neglected. For example, although Jacob et al. [4] uses transformer, it only imposes constraints on convolutional attention.

Another line of solution involves modeling the relationships among AUs. There are efforts in proposing adaptive spatio-temporal graph convolutional networks to reason each AU's independent pattern, and employing relation learning layers to capture different AU relations [6]. However, these approaches ignore the modeling of long-range dependencies among facial regions. Long-range dependencies denote that some facial distant AUs have positive correlations (co-occurrence) or negative correlations (mutual exclusion). For example, surprised expression often has the co-occurrences of AU 1, AU 2, AU 5, and AU 26, where AU 1 and AU 2 are in the eyebrow region, and AU 26 is in the chin region.

To tackle the above issues, we propose an **E**nsemble **S**elf-attention **C**onstraining and **D**ual-directional **A**ttention (**ESCDA**) method, which combines a local self-attention constraining (LSC) network to model self-attention distribution and a global dual-directional attention (GDA) network to capture long-range relationships. Specifically, we enhance the feature extraction capacity of MobileFaceNet [7] by integrating coordinate attention [8] and mixed depthwise convolution (MixConv) [9] as the backbone. In LSC, we develop a self-attention constraining head for each AU-specific branch, by regarding the self-attention distribution as a spatial distribution, and constraining it via prior knowledge so as to capture AU-related local information while long-range relational modeling ability can be preserved. Furthermore, in GDA, we design a spatial dual-directional attention head, which adaptively learns global attention map from both vertical and horizontal directions. By exploiting dual-directional information flows within the spatial domain, the simultaneous prediction of multiple AUs is facilitated. This method can capture not only long-distance dependencies between regions but also image's global context.

Additionally, we observe that recent works tend to train deeper and more complex networks. However, there have been few works explored effective ensemble methods for AU detection task. Allen-Zhu et al. [10] confirmed through experiments and theoretical analysis that randomly initialized neural networks may focus on features from different perspectives. Consequently, ensemble models, which can simultaneously attend to multi-view features, generally achieve improved performance. Inspired by Allen-Zhu et al. [10], we adopt a multi-view ensemble method to integrate the strengths of both models. However, instead of using randomly initialized neural networks, two models from different views specifically designed for AU detection were employed. In particular, we train both models in parallel and subsequently average their label predictions. Multi-view ensemble enables our method to more comprehensively capture data characteristics from both local and global area [11], and thus improves the generalization ability.

The main contributions of this work include:

- We propose a novel local self-attention constraining network, which constrains self-attention distribution based on prior location knowledge to effectively capture local information associated with AUs.
- We propose a novel global dual-directional attention network, which captures the correlations among multiple AUs across the face by learning attentions from two directions.
- We introduce a multi-view ensemble method to merge the strengths of both models, boosting their abilities to extract both local and global features.
- Extensive experimental results demonstrate that our methods including single models and ensemble model all achieve state-of-the-art performance on challenging BP4D, DISFA, and GFT datasets, particularly outperforming previous works significantly on DISFA and GFT.

2. Related work

We review the previous approaches that are relevant to our work, including self-attention based AU detection, regional learning based AU Detection, and ensemble learning in computer vision.

2.1. Self-attention based AU detection

In recent years, vision transformers with self-attention mechanism have revolutionized the field of computer vision. Since the self-attention can effectively capture long-range dependencies and learn rich contextual information, a few works have incorporated it to precisely detect subtle AUs. Yuan et al. [12] used a frozen pre-trained Vision Transformer combined with lightweight modules for efficient feature learning. Li et al. [13] introduced the self-attention into AU detection task, which is learned through the AU label supervision. Jacob et al. [4] captured the relationship between different AUs via self-attention. However, the distribution of self-attention weights has semantic characteristic, and such underlying mechanism has been overlooked.

In contrast, we propose to regard the self-attention distribution of each AU as a spatial pattern, and exploit prior knowledge about AU locations to constrain the learning of self-attention.

2.2. Regional learning based AU detection

AU refers to facial local muscle movements, and thus extracting its features requires accurately locating associated areas. Considering the challenge of capturing AU features, Jaiswal et al. [14] used convolutional neural networks (CNNs) to extract features of each AU from cropped regions of interest (ROIs) and masks. Zhao et al. [15] implemented convolution layers that consist of multiple independent blocks, where each block contains its own convolution filters to extract features. Ma et al. [16] defined bounding boxes for AUs using landmarks and incorporated general object detection backbone into AU detection. Shao et al. [17] adaptively learned channel-level and spatial attentions from AU detection supervisions, while Liu et al. [18] adaptively updated AU correlation graphs by efficiently leveraging multi-level AU motion and emotion features extracted at different network stages. Ge et al. [19] performed region-level, pixel-level, and channel-level feature

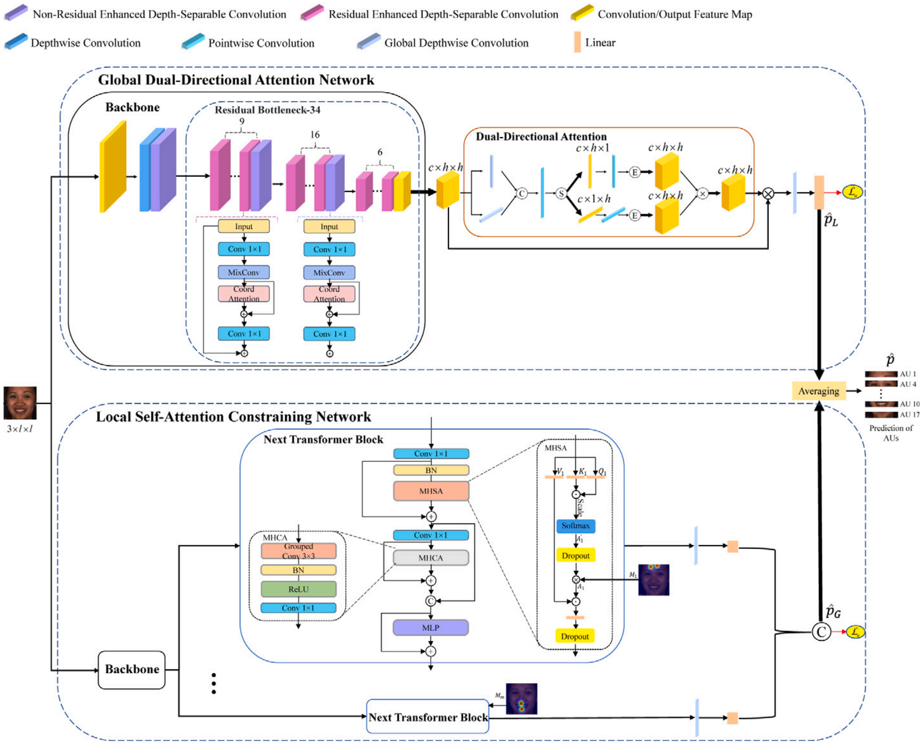


Fig. 2. The structure of our ESCDA framework (zoom in for a better view). An input image is initially processed by the single-branch global dual-directional attention network and the m-branch local self-attention constraining network. Both networks have the same backbone structure, the same AU detection loss $\mathcal{L}_u$, and similar AU classifier structure composed of a global depthwise convolutional layer and a linear layer. The outputs from both, $\hat{\mathbf{p}}_L$ and $\hat{\mathbf{p}}_G$, are averaged to produce the final prediction $\hat{\mathbf{p}}$. "⊗" and "⊕" denote element-wise multiplication and addition, respectively; "⊙" denotes matrix multiplication; "C", "S", and "E" in a circle denote concatenation, split, and expansion, respectively.

learning. Chen et al. [20] embedded 3D manifold information into 2D convolutions. However, convolution based methods struggle to capture long-range AU information, and positional information, which can help the model focus on the key regions of input images, is often neglected by channel attention based methods.

Different from these works, we propose the global dual-directional attention network to learn global attentions adaptively from both vertical and horizontal directions, so as to capture long-distance dependencies among facial regions.

2.3. *Ensemble learning in computer vision*

Ensemble learning is a technique that seeks better prediction performance by combining the predictions from multiple models. It has been introduced to the field of computer vision. For example, Moghimi et al. [21] proposed a method that merges boosting with deep learning, which utilizes the least squares objective function. Li et al. [22] developed a sparse deep stacking network for image classification, which employed mixed-norm regularization to learn sparse representations.

Recently, ensemble learning has gained attention in AU detection. Jiang et al. [23] combined models according to the highest F1 scores for each AU. Jeong et al. [24] employed the Bagging method by training each classifier with a subset of the data and utilizing soft voting to amalgamate predictions from various models. Although these works utilize ensemble learning to boost model performance, they do not provide a detailed justification for its effectiveness. AAR [6], composed of an adaptive attention regression network and an adaptive spatio-temporal graph convolutional network, is equivalent to the stacking method of ensemble learning. Our method is to calculate the average output of two not-too-complex networks, but the effect is significantly improved. This may inspire subsequent works to focus on multi-perspective ensemble.

3. Methodology

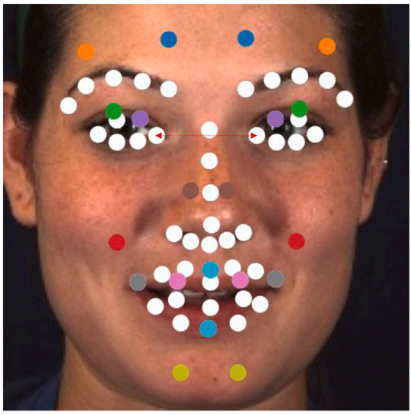
Given an image with the size of $3 \times l \times l$, our main goal is to estimate the AU occurrence probabilities $\hat{\mathbf{p}} = (\hat{p}^{(1)}, \ldots, \hat{p}^{(m)})$, in which m is the number of evaluated AUs. The architecture of our ESCDA framework is shown in Fig. 2. Initially, the image is fed into the local self-attention constraining (LSC) network and the global dual-directional attention (GDA) network to extract facial information from local and global perspective, respectively. Since model ensemble can significantly enhance AU detection performance by learning features from various perspectives, the outputs from these networks are further averaged to derive the final prediction.

3.1. *Local self-attention constraining*

The structure of LSC network is illustrated in the bottom part of Fig. 2. Since the next transformer block (NTB) [25] achieves a better trade-off between latency and accuracy, we adopt it as the self-attention module and apply self-attention constraints to it. The input image first goes through the backbone for feature extraction, then m branches are followed. Each branch comprises a next transformer block, a global depthwise convolutional layer, and a linear layer to identify a certain AU in local regions.

3.1.1. *Backbone*

Our backbone is developed from MobileFaceNet [7], which is a lightweight network with superior performance in face detection task. In particular, the input first passes through a convolutional layer to extract preliminary features, and then proceeds to depthwise convolution and enhanced depth-separable convolution for shallow feature extraction. Within the enhanced depth-separable convolution, MixConv [9] and coordinate attention [8] are introduced. MixConv enhances feature



AU	Location
● 1 (Inner brow raiser)	1/2 scale above inner brow
● 2 (Outer brow raiser)	1/3 scale above outer brow
● 4 (Brow lowerer)	1/3 scale below brow center
● 6 (Cheek raiser)	1 scale below eye bottom
● 7 (Lid tightener)	Eye
● 9 (Nose wrinkler)	1/2 scale above nose bottom
● 10 (Upper lip raiser)	Upper lip center
12 (Lip corner puller)	Lip corner
● 14 (Dimpler)	
15 (Lip corner depressor)	
17 (Chin raiser)	1/2 scale below lip
● 26 (Jaw drop)	
23 (Lip tightener)	Lip center
● 24 (Lip pressor)	
25 (Lips part)	

Fig. 3. The definition and visualization for the positions of AU centers, which are designed for a face aligned such that two eye centers are horizontal [5,26]. For each AU, two centers are determined based on two associated facial landmarks. A red dotted line is used to indicate "scale", representing the distance between the inner corners of two eyes.

representation by integrating multi-scale convolution, whereas coordinate attention boosts the modeling of long-range dependencies. Following shallow feature extraction, deeper feature extraction is achieved by stacking two types of enhanced depth-separable convolutions. We employ three levels of residual and non-residual enhanced depth-separable convolutions, to downsample the feature map sizes successively to 1/2, 1/4, and 1/8, while simultaneously increasing the number of channels.

3.1.2. Adaptive constraining on self-attention distribution

To address the shortcomings of self-attention in capturing local features, we propose to apply self-attention constraining to the NTB to assist LSC network in capturing local feature dependencies. As illustrated in Fig. 3 , correlated landmarks can accurately locate the central area of each AU based on prior positional relationships [26]. To leverage such prior knowledge, we predefine a mask $\mathbf{M}_i$ for the ith AU, which has two highlighted sub-centers $(\bar{a}_{i(1)}, \bar{b}_{i(1)})$ and $(\bar{a}_{i(2)}, \bar{b}_{i(2)})$. We initially create the predefined mask $\widetilde{\mathbf{M}}_{i(1)}$ for one sub-center $(\bar{a}_{i(1)}, \bar{b}_{i(1)})$ using a Gaussian distribution with standard deviation δ. The value at location (a, b) is defined as

$$\widetilde{M}_{iab(1)} = \exp(-\frac{(a-\bar{a}_{i(1)})^2 + (b-\bar{b}_{i(1)})^2}{2\delta^2}). \tag{1}$$

Then, we combine $\widetilde{\mathbf{M}}_{i(1)}$ and $\widetilde{\mathbf{M}}_{i(2)}$ by selecting the maximum value at each location (a, b):

$$\widetilde{M}_{iab} = \max(\widetilde{M}_{iab(1)}, \widetilde{M}_{iab(2)}) \in (0, 1]. \tag{2}$$

In $\widetilde{\mathbf{M}}_i$, the region of interest (ROI) for the ith AU comprises positions with values greater than 0, while disregarding other positions with zero values. However, this could potentially result in losing some valuable underlying information. To give a certain degree of importance to areas outside the ROI, we introduce a learnable parameter ϵ_i:

$$M_{iab} = \frac{\widetilde{M}_{iab} + \epsilon_i}{1 + \epsilon_i} \in (0, 1], \tag{3}$$

where $\epsilon_i \geq 0$, and is proportional to the importance of the area outside the ROI.

On the other hand, the scaled dot-product attention weight in the self-attention mechanism [27] is calculated as

$$\mathbf{A}'_i = Softmax(\frac{\mathbf{Q}_i \mathbf{K}_i^T}{\sqrt{d'}}), \tag{4}$$

where $\mathbf{Q}_i \in \mathbb{R}^{k' \times n' \times d'}$, $\mathbf{K}_i \in \mathbb{R}^{k' \times n' \times d'}$, $\mathbf{A}'_i \in \mathbb{R}^{k' \times n' \times n'}$, and Softmax function is denoted as $Softmax(\cdot)$. Afterward, we derive the constrained scaled dot-product attention weight by element-wise multiplying the obtained predefined mask $\mathbf{M}_i$ with $\mathbf{A}'_i$:

$$\mathbf{A}_i = \mathbf{A}'_i \otimes \mathbf{M}_i, \tag{5}$$

in which $\otimes$ denotes elementwise multiplication. Constrained $\mathbf{A}_i$ has multiple attention channels to derive features from AU-related regions. Then, the self-attention process is defined as

$$SA(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i) = \mathbf{A}_i \mathbf{V}_i, \tag{6}$$

where $\mathbf{V}_i \in \mathbb{R}^{k' \times n' \times d'}$.

We further combine the self-attention $SA(\cdot)$ from multiple heads by multiplying by a learnable weight matrix $\mathbf{W}^M$ to create the multi-head self-attention (MHSA). Afterward, a pointwise convolutional layer and a multi-head convolutional attention (MHCA) [25] block are followed, in which the output is concatenated with the output of MHSA to fuse information. Finally, a multi-layer perceptron (MLP) layer is employed for further feature learning.

3.1.3. AU detection

After NTB, we use an AU classifier composed of a global depthwise convolutional layer and a linear layer in each AU branch to obtain the predicted AU occurrence probability $\hat{\mathbf{p}}_L$. The weighted AU detection loss [6] we use is defined as

$$\mathcal{L}_u = -\sum_{i=1}^{m} w^{(i)} [v^{(i)} p^{(i)} \log \hat{p}_L^{(i)} + (1 - p^{(i)}) \log(1 - \hat{p}_L^{(i)})], \tag{7}$$

in which $w^{(i)}$, $v^{(i)}$, and $p^{(i)}$ denote the weight, occurrence weight, and the ground-truth occurrence probability of the i-AU, respectively, and $p^{(i)}$ is 1 for occurrence or 0 for non-occurrence. Most AU datasets [28, 29] exhibit two types of data imbalance issues: certain AUs have higher occurrence rates compared to others, and each AU has a higher non-occurrence rate than occurrence rate. To alleviate the issues associated with data imbalance, $w^{(i)}$ and $v^{(i)}$ are defined as

$$w^{(i)} = \frac{n}{n_{occ}^{(i)}} / \sum_{k=1}^{m} \frac{n}{n_{occ}^{(k)}}, \quad v^{(i)} = \frac{n - n_{occ}^{(i)}}{n_{occ}^{(i)}}, \tag{8}$$

where $n_{occ}^{(i)}$ represents the number of occurrences of the ith AU, while n denotes the total number of images in the training dataset.

3.2. Global dual-directional attention

To capture the relationships and interactions between global area of AUs, channel attention is divided into two one-dimensional feature coding processes, which aggregate features along two spatial directions, allowing to get remote dependencies in one spatial direction while maintaining position information in the other spatial direction. The linear global depthwise convolution adopted in GDA can extract detailed features from various areas of the face effectively. As a result, our GDA network establishes long-range dependencies between different AU areas through the dual-directional attention mechanism, realizing the fusion and propagating of AU information on a global scale. As illustrated in the upper part of Fig. 2, the input is passed through the backbone to generate a feature map, which is then multiplied by the attention map generated through the dual-directional attention module. Finally, the prediction result is obtained through an AU classifier. Except for a single AU classifier to predict all AUs, the backbone, AU classifier, and AU detection loss $\mathcal{L}_u$ within the GDA network are identical to their counterparts in the LSC network.

Since coordinate attention [8] simultaneously captures long-distance dependencies in both the height and width directions, we incorporate it into the dual-directional attention module. Initially, we transform the input feature map in both vertical and horizontal directions to extract directional features. Note that we use global depthwise convolution instead of average pooling. This is because global depthwise convolution can adaptively extract global features from feature maps using learnable convolution kernels, enabling the learning of a more efficient global feature representation. After concatenating vertical and horizontal features and processing them through pointwise convolution, they are split, and pointwise convolution is then applied to generate vertical and horizontal attention maps. Finally, we perform an element-wise multiplication of the expanded attention maps to obtain a new attention map. This new attention map will be multiplied with the feature map to adaptively learn the global AU relationship.

3.3. Multi-view ensemble

When learning multi-view features, there are alignment issues due to differences between the global and local features of AUs. Ensemble learning allows the final model to learn features from different perspectives. Allen-Zhu et al. [10] offered a theoretical foundation that models trained independently through averaging can effectively capture multi-view features. Such multi-view ensemble contributes to the target variable, and thus enhances the test accuracy.

Specifically, given data X and label y from the training set $\mathcal{Z}$, we have $K = \tilde{\Omega}(1)$ models $\{F^{[\ell]}\}_{\ell\in[K]}$, independently trained for $T = O(\frac{poly(k)}{\eta})$ iterations. The ensemble model G is defined as

$$G(X) = \frac{\tilde{\Theta}(1)}{K} \sum_{\ell} F^{[\ell]}(X). \tag{9}$$

Here, $\tilde{\Theta}(1)$ represents constant time complexity, and the entire ensemble process essentially is averaging the outputs of a few independently trained models. Then, in training set, we have the following formulations with probability at least $1 - e^{-\Omega(\log^2 k)}$:

$$(X, y) \in \mathcal{Z}, i \in [k]\backslash\{y\} : G_y(X) > G_i(X), \tag{10a}$$

that is, for each sample in the training data, the score of the correct category exceeds those of all other categories. And in test set, we have:

$$\Pr_{(X,y)\sim D}\left[\exists i \in [k]\backslash\{y\} : G_y(X) < G_i(X)\right] \le 0.001\mu. \tag{10b}$$

This indicates that under the test distribution D, the probability of misclassification is strictly controlled at a very low level, thereby ensuring the model's generalization performance. The detailed proof of this theorem can be seen in [10].

Not only does Eq. (10a) ensure that the ensemble model guarantees correct classification on the training set, but also Eq. (10b) ensures that the ensemble model achieves high generalization performance on the test set. Additionally, the ensemble requires only a small number $K = \tilde{\Omega}(1)$ of individually trained models. In Allen-Zhu et al.'s work [10], the fundamental reason for the validity of the theorem lies in the fact that these "lottery winning sets" $\mathcal{M}$, which are the appropriately initialized or selected models, enabling to achieve high performance, are generated based on the random initialization of the neural network.

AU detection not only requires attention to local muscle movements but also necessitates integrating the relationships among global regions. In our work, unlike Allen-Zhu et al. [10] randomly initializing neural networks to generate lottery winning sets, we intentionally designed two models from local and global views to accommodate different lottery winning sets for the AU task. The LSC primarily correspond to the local features of the face, that is, those subtle yet discriminative local variations; Whereas the GDA encompass the global structural information of the face and the interrelationships between different regions. When we employ the ensemble learning averaging strategy to fuse the outputs of these two models, it essentially takes the union of the two winning sets, ensuring that all important features are sufficiently captured.

After we obtain the prediction $\hat{\mathbf{p}}_L$ of the LSC network and the prediction $\hat{\mathbf{p}}_G$ of the GDA network, the final AU occurrence probabilities $\hat{\mathbf{p}}$ is calculated as $\hat{\mathbf{p}} = (\hat{\mathbf{p}}_L + \hat{\mathbf{p}}_G)/2$.

4. Experiments

4.1. Datasets and settings

4.1.1. Datasets

Our ESCDA is evaluated on three benchmark datasets, in terms of BP4D [28], DISFA [29], and GFT [30].

- **BP4D** comprises 23 female and 18 male participants, each involved in 8 sessions. It contains approximately 140,000 frames, each annotated for the occurrence or non-occurrence of specific AUs, in addition to 49 facial landmarks. Our evaluation, adhering to settings in [26], focuses on 12 AUs (1, 2, 4, 6, 7, 10, 12, 14, 15, 17, 23, and 24) through a subject-exclusive 3-fold cross-validation process, allocating two folds for training and one for testing.
- **DISFA** contains 27 videos, captured from 12 females and 15 males, each including 4,845 frames. Each frame is annotated with AU intensities on a six-point ordinal scale ranging from 0 to 5, and includes 66 facial landmarks. This dataset has a more severe data imbalance problem than BP4D dataset.
- **GFT** encompasses 96 subjects divided into 32 three-subject groups, engaging in unscripted conversations. Each subject is recorded in a video featuring mostly moderate out-of-plane poses. Each video frame is annotated with 10 AUs (1, 2, 4, 6, 10, 12, 14, 15, 23, and 24) and 49 facial landmarks. We follow the official partitions [30], using 78 subjects with around 108,000 frames for training, and 18 subjects with around 24,600 frames for testing.

4.1.2. Implementation details

Each face image is aligned to 3 × 200 × 200 size through similarity transformation based on facial landmarks. For training data augmentation, images are randomly cropped to 3 × 176 × 176, resized to 3 × 112 × 112 for input into our networks, and subjected to random mirroring and color jittering for contrast and brightness adjustments. The crop size l, the dimension parameters c and h, and the standard deviation δ are set as 112, 512, 7, and 3, respectively. The number of AUs m is 12 for BP4D, 8 for DISFA, and 10 for GFT.

We implement our LSC and GDA using PyTorch. The backbone networks of both LSC and GDA are pretrained on MS-Celeb-1M dataset [31]. Then, both networks are trained using identical experimental configurations, including being trained for up to 20 epochs using the AdamW optimizer, employing a cosine decay learning rate scheduler, a 5-epoch linear warm-up, an initial learning rate of $3.2 \times 10^{-2}/256$ multiplying the mini-batch size, a weight decay of 0.05, and gradient clipping with a max norm of 3.0. LSC and GDA can be simultaneously trained, in which we select the best average predictions as the prediction results of our ensemble model ESCDA.

Table 1
F1-frame results for 12 evaluated AUs on BP4D [28]. The results of previous methods are reported in original papers.

AU	1	2	4	6	7	10	12	14	15	17	23	24	**Avg**
DRML [15]	36.4	41.8	43.0	55.0	67.0	66.3	65.8	54.1	33.2	48.0	31.7	30.0	48.3
EAC-Net [5]	39.0	35.2	48.6	76.1	72.9	81.9	86.2	58.8	37.5	59.1	35.9	35.8	55.9
ARL [17]	45.8	39.8	55.1	75.7	77.2	82.3	86.6	58.8	47.6	62.1	47.4	55.4	61.1
JÂA-Net [26]	53.8	47.8	58.2	78.5	75.8	82.7	88.2	63.7	43.3	61.8	45.6	49.9	62.4
AU R-CNN [16]	50.2	43.7	57.0	78.5	78.5	82.6	87.0	67.7	49.1	62.4	50.4	49.3	63.0
AAR [6]	53.2	47.7	56.7	75.9	79.1	82.9	88.6	60.5	51.5	61.9	51.0	56.8	63.8
Jacob et al. [4]	51.7	**49.3**	61.0	77.8	79.5	82.9	86.3	67.6	51.9	63.0	43.7	56.3	64.2
CISNet [32]	54.8	48.3	57.2	76.2	76.5	**85.2**	87.2	66.2	50.9	65.0	47.7	56.5	64.3
Chang et al. [33]	53.3	47.4	56.2	79.4	80.7	85.1	**89.0**	67.4	55.9	61.9	48.5	49.0	64.5
KS [34]	55.3	48.6	57.1	77.5	**81.8**	83.3	86.4	62.8	52.3	61.3	51.6	**58.3**	64.7
SMA-ViT [35]	52.7	45.6	59.8	**83.8**	79.2	83.5	87.2	64.0	54.1	61.2	**52.6**	**58.3**	65.2
AUNet [36]	**58.0**	48.2	**62.4**	76.4	77.5	83.4	88.5	63.3	52.0	**65.5**	52.1	52.3	65.0
MGRR-Net [19]	52.6	47.9	57.3	78.5	77.6	84.9	88.4	**67.8**	47.6	63.3	47.4	51.3	63.7
Liu et al. [18]	57.8	48.8	59.4	79.1	78.8	84.0	88.2	65.2	56.1	63.8	50.8	55.2	**65.6**
LSC	54.9	48.8	61.3	77.2	76.8	84.4	86.9	58.6	53.4	65.1	50.7	51.9	64.2
GDA	53.3	42.3	58.6	78.3	76.6	83.9	88.4	65.0	54.5	62.2	51.0	56.4	64.2
ESCDA	56.5	44.5	59.6	79.4	77.4	84.7	88.7	64.9	**57.0**	65.0	52.5	55.6	65.5

Table 2
F1-frame results for 8 evaluated AUs on DISFA [29]. The results of previous methods are reported in original papers.

AU	1	2	4	6	9	12	25	26	Avg
DRML [15]	17.3	17.7	37.4	29.0	10.7	37.7	38.5	20.1	26.7
EAC-Net [5]	41.5	26.4	66.4	50.7	8.5	**89.3**	88.9	15.6	48.5
AU R-CNN [16]	32.1	25.9	59.8	55.3	39.8	67.7	77.4	52.6	51.3
ARL [17]	43.9	42.1	63.6	41.8	40.0	76.2	95.2	66.8	58.7
SMA-ViT [35]	51.2	49.3	64.7	48.3	50.6	87.6	85.1	61.2	62.2
KS [34]	53.8	59.9	69.2	54.2	50.8	75.8	92.2	46.8	62.8
JÂA-Net [26]	62.4	60.7	67.1	41.1	45.1	73.5	90.9	67.4	63.5
AAR [6]	62.4	53.6	71.5	39.0	48.8	76.1	91.3	70.6	64.2
Chang et al. [33]	60.4	59.2	67.5	52.7	51.5	76.1	91.3	57.7	64.5
CISNet [32]	48.8	50.4	**78.9**	51.9	47.1	80.1	95.4	65.0	64.7
AUNet [36]	60.3	59.1	69.8	48.4	53.0	79.7	93.5	64.7	66.1
MGRR-Net [19]	61.3	62.9	75.8	48.7	53.8	75.5	94.3	**73.1**	68.2
Liu et al. [18]	62.0	65.7	74.5	53.2	43.1	76.9	**95.6**	53.1	65.5
LSC	67.1	62.4	68.8	51.3	51.8	75.6	94.9	60.2	66.5
GDA	**68.1**	62.3	70.4	52.8	**55.4**	76.4	93.5	67.6	68.3
ESCDA	65.3	**66.7**	72.1	**56.4**	55.1	77.6	94.9	68.6	**69.6**

Table 3
F1-frame results for 10 evaluated AUs on GFT [30]. The results of EAC-Net [5] and ARL [17] are reported in [26].

AU	1	2	4	6	10	12	14	15	23	24	Avg
EAC-Net [5]	15.5	56.6	0.1	81.0	76.1	84.0	0.1	38.5	57.8	51.2	46.1
TCAE [37]	43.9	49.5	6.3	71.0	76.2	79.5	10.7	28.5	34.5	41.7	44.2
ARL [17]	51.9	45.9	13.7	79.2	75.5	82.8	0.1	44.9	59.2	47.5	50.1
Ertugrul et al. [38]	43.7	44.9	19.8	74.6	76.5	79.8	**50.0**	33.9	16.8	12.9	45.3
JÂA-Net [26]	46.5	49.3	19.2	79.0	75.0	84.8	44.1	33.5	54.9	50.7	53.7
AAR [6]	**66.3**	53.9	23.7	81.5	73.6	84.2	43.8	**53.8**	58.2	46.5	58.5
LSC	66.2	58.7	49.4	85.9	77.4	84.8	30.0	52.6	59.7	53.0	61.8
GDA	62.7	57.8	**62.4**	84.8	**81.1**	85.4	29.0	51.3	58.5	**53.5**	62.7
ESCDA	65.5	**58.9**	54.6	**87.0**	79.6	**86.2**	31.7	53.7	**60.5**	52.7	**63.0**

4.1.3. Evaluation metrics

We utilize a widely used metric, known as frame-based F1-score (F1-frame). It is defined as $F1 = 2PR/(P+R)$, where P stands for precision, and R represents recall. Additionally, we present the average F1-frame results across all AUs, abbreviated as Avg. The F1-frame results for subsequent sections are all expressed in percentages, albeit without the "%" symbol.

4.2. Comparison with state-of-the-art methods

Our ESCDA is compared against state-of-the-art AU detection methods under the same evaluation setting, including DRML [15], EAC-Net [5], ARL [17], AU R-CNN [16], JÂA-Net [26], Jacob et al. [4], AAR [6], CISNet [32], Chang et al. [33], KS [34], SMA-ViT [35], TCAE [37], Ertugrul et al. [38], AU-Net [36], MGRR-Net [19], and Liu et al. [18]. All these methods are based on deep learning, most of which are relevant to our approach in terms of self-attention and regional learning. Table 1 and Table 2 present the results for individual AUs and the overall average results. We can see that our technique achieves consistently better results than state-of-the-art methods on BP4D, DISFA and GFT, as indicated by higher average F1-frame.

4.2.1. Evaluation on BP4D

As demonstrated in Table 1, without model ensemble, our proposed LSC and GDA methods already outperform most state-of-the-art approaches. For instance, LSC achieves higher average F1-frame than recent self-attention based work Jacob et al. and GDA achieves higher average F1-frame than recent regional learning based works such as JÂA-Net and AAR. These results highlight the effectiveness of our method in capturing local AU features and global AU relationships, respectively. After assembling the two models, our ESCDA outperforms all other methods, though it performs slightly worse than Liu et al., and the performance across AUs is more balanced than most methods. This validates the effectiveness of multi-view ensemble, in which our ensemble model can effectively capture both local-view features from local AU areas and global-view features of global AU relationships by averaging the predictions. Our method is a simple but effective solution to the challenging AU detection task.

4.2.2. Evaluation on DISFA

As shown in Table 2, our LSC and GDA both significantly outperform all existing methods, and the performance of our ensemble model ESCDA is further improved, achieving a 1.4 margin in terms of average F1-frame over the best existing method MGRR-Net. Note that the data imbalance issue in the DISFA dataset is more severe compared to that of BP4D, leading to performance fluctuations in previous works such as AU-GCN. In this challenging case, our models including LSC, GDA, and ESCDA all exhibit strong and stable performance among different AUs. This can be attributed to our proposed local self-attention constraining, global dual-directional attention, and multi-view ensemble. Especially, multi-view ensemble is beneficial for learning more robust and comprehensive features, enabling our ESCDA to effectively handle the challenges posed by data imbalance.

4.2.3. Evaluation on GFT

Table 3 presents the F1-frame results for the GFT benchmark. It can be seen that our methods including LSC, GDA, and the ensemble model ESCDA all significantly outperform previous works. Compared to BP4D and DISFA, GFT images display moderate deviations from the frontal plane. In this challenging case, ESCDA achieves strong performance with an average F1-frame score of 63.0.

Table 4
Floating point operations (FLOPs) and the number of parameters (#Params.) for different methods during the detection of 12 AUs.

Method	FLOPs	#Params.
DRML [15]	**0.9G**	56.9M
EAC-Net [5]	18.8G	337.5M
JÂA-Net [26]	8.8G	25.2M
AAR [6]	10.2G	7.2M
CISNet [32]	4.8G	22.4M
GDA	4.5G	**4.1M**
LSC	13.7G	27.2M

Table 5
F1-frame results for 12 evaluated AUs from different variants of ESCDA on BP4D [28]. The best results are highlighted in bold, and the second best results are highlighted by an underline.

AU	1	2	4	6	7	10	12	14	15	17	23	24	Avg
LS	55.6	45.7	56.1	76.6	74.6	82.7	88.1	61.0	54.3	**65.3**	49.9	50.6	63.4
LSC$^{(fix)}$	51.7	44.9	59.4	77.2	76.3	82.8	87.1	61.5	55.8	62.4	**53.2**	56.2	64.0
LSC	54.9	**48.8**	61.3	77.2	76.8	84.4	86.9	58.6	53.4	65.1	50.7	51.9	64.2
UA	51.7	48.0	**61.7**	78.0	76.4	83.1	88.3	58.3	54.8	63.5	48.4	52.6	63.7
GDA	53.3	42.3	58.6	78.3	76.6	83.9	88.4	65.0	54.5	62.2	51.0	56.4	64.2
ESCDA-V	52.1	43.4	60.6	78.4	**78.5**	84.6	88.5	**66.5**	55.4	63.1	50.9	**56.9**	64.9
ESCDA	**56.5**	44.5	59.6	**79.4**	77.4	**84.7**	**88.7**	64.9	**57.0**	65.0	52.5	55.6	**65.5**

Table 6
The structures of different variants of our ESCDA. **B**: enhanced MobileFaceNet backbone. **N**: vanilla NTB. **N^c**: self-attention constraining in NTB. $\mathbf{M}^{(fix)}$: predefined mask $\mathbf{M}_i$ with fixed $\epsilon_i = 0$ for the i-th AU. $\mathbf{M}^{(ada)}$: predefined mask $\mathbf{M}_i$ with adaptively learned ϵ_i for the i-th AU. $\mathbf{A^u}$: unidirectional attention in global face. $\mathbf{A^d}$: dual-directional attention in global face. **C**: AU classifier. $\mathbf{E^v}$: model ensemble using the voting method. $\mathbf{E^a}$: model ensemble using the averaging method.

Method	**B**	**N**	$\mathbf{N^c}$	$\mathbf{M^{(fix)}}$	$\mathbf{M^{(ada)}}$	$\mathbf{A^u}$	$\mathbf{A^d}$	**C**	$\mathcal{L}_u$	$\mathbf{E^v}$	$\mathbf{E^a}$
LS	√	√						√	√		
LSC$^{(fix)}$	√		√	√				√	√		
LSC	√		√		√			√	√		
UA	√					√		√	√		
GDA	√						√	√	√		
ESCDA-V	√		√		√		√	√	√	√	
ESCDA	√		√		√		√	√	√		√

Table 7
F1-frame results for 8 evaluated AUs from different variants of ESCDA on DISFA [29]. The best results are highlighted in bold, and the second best results are highlighted by an underline.

AU	1	2	4	6	9	12	25	26	Avg
LS	64.6	54.6	58.8	54.3	37.3	**77.7**	**96.2**	63.5	63.4
LSC$^{(fix)}$	60.8	63.1	65.1	56.1	48.9	76.2	94.9	52.7	64.7
LSC	67.1	62.4	68.8	51.3	51.8	75.6	94.9	60.2	66.5
UA	67.0	64.4	67.9	48.7	48.5	76.5	92.8	65.6	66.4
GDA	**68.1**	62.3	70.4	52.8	**55.4**	76.4	93.5	67.6	68.3
ESCDA-V	67.9	62.7	66.6	54.4	47.7	77.1	92.6	66.1	66.9
ESCDA	65.3	**66.7**	**72.1**	**56.4**	55.1	77.6	94.9	**68.6**	**69.6**

4.2.4. Discussion about model complexity

In Table 4, the floating point operations (FLOPs) and the number of parameters (#Params.) of different methods for 12 AUs are shown. Many existing methods do not release their code or report metrics such as FLOPs and the number of parameters. Therefore, we compare our approach only with those methods that have made their code or model complexity publicly available. As we can see, GDA has the fewest model parameters compared to previous methods, and its FLOPs are the second smallest, only slightly higher than DRML [15]. Despite this, GDA delivers significant performance improvements, demonstrating that the dual-directional attention module offers both low computational cost and high efficiency. LSC has more parameters and FLOPs comparable to earlier works like JÂA-Net [26] and CISNet [32]. Inevitably, its multi-branch structure and next transformer blocks lead to increased memory and time costs.

However, we believe that sacrificing computation for performance improvements is worthwhile, as AU detection is a micro-action-sensitive task where enhancing accuracy is both challenging and crucial [6]. Previous works like Deng et al. [39] proposed a multimodal fusion framework that integrates visual and audio features, utilizing various deep learning models for feature integration. Although the model needs to handle multimodal inputs and long-sequence training, its performance significantly outperforms that of unimodal approaches. Additionally, the computational overhead can be mitigated by advancements in hardware, and processing can be handled on cloud servers with results delivered to clients. In the future, we plan to reduce the model's complexity by incorporating more efficient networks and optimizing the multi-branch architecture.

4.3. Ablation study

In this section, we explore the significance of each key component within our ESCDA framework. The F1-frame results for different ESCDA variants on BP4D are displayed in Table 5, with the structure of each variant detailed in Table 6.

4.3.1. Adaptive constraining on self-attention distribution

The baseline LS creates a branch containing an NTB for each AU. Furthermore, LSC$^{(fix)}$ and LSC apply self-attention constraining within NTB's MHSA module. Compared to LSC$^{(fix)}$, LSC introduces an adaptively learnable parameter ϵ_i for each AU branch, enabling more flexible identification of regions beneficial for AU detection. We can observe a gradual increase in the average F1-frame from 63.4 to 64.0 and then to 64.2, demonstrating the effectiveness of our approach, which treats the self-attention distribution of each AU as a spatial distribution and adaptively imposes constraints based on prior knowledge.

4.3.2. Global dual-directional attention

Another baseline UA only learns unidirectional attention in global face. Based on UA, GDA uses the full dual-directional attention module, enabling the adaptive learning of global attention map in both horizontal and vertical directions. This results in the average F1-frame improving from 63.7 to 64.2, demonstrating GDA's ability to understand the relationships among global AUs through dual-directional attention.

4.3.3. Multi-view ensemble

We notice that the average F1-frame of LSC is close to that of GDA. Combining LSC and GDA into an ensemble, the model's average F1-frame rose from 64.2 to 65.5. This notable improvement underscores multi-view ensemble's efficacy. LSC captures local patterns, while GDA captures global ones. Through multi-view ensemble, our ESCDA combines these models' strengths to detect various pattern types. Concurrently, we can see that ensemble learning by the voting method also improves model performance, though not as effectively as the averaging method. This is because taking the maximum value of two models in the voting method may lose valuable information and overlooking more precise predictions, whereas averaging focuses on diverse patterns, facilitating a more effective integration of both models' strengths.

4.3.4. More discussions on DISFA

Since each of our models achieves remarkable performance on the DISFA dataset, we also conduct ablation studies to further investigate their effectiveness. As observed in Table 7, ESCDA achieves either the best or second-best results for all AUs except AU 1, demonstrating that ensemble learning can effectively combine the strengths of LSC

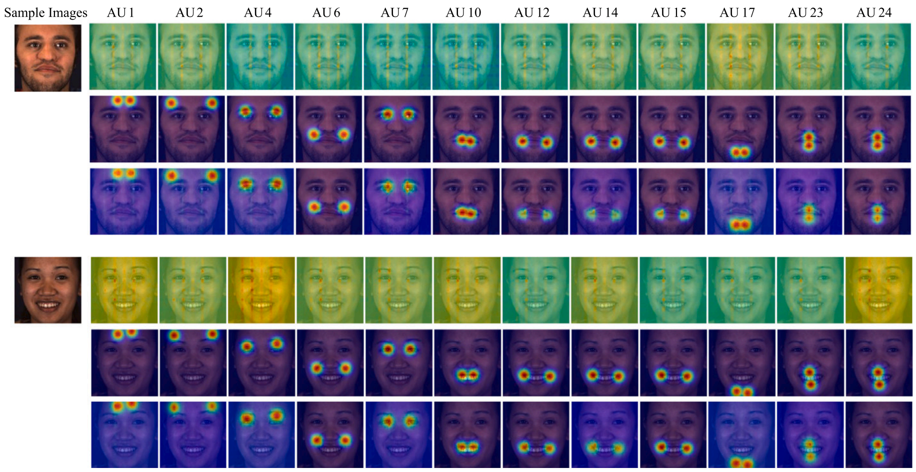


Fig. 4. Visualization of pre-constrained self-attention $\mathbf{A}'_i$ averaged over channels, predefined mask $\mathbf{M}_i$, post-constrained self-attention $\mathbf{A}_i$ averaged over channels learned by our LSC on two sample images from BP4D [28]. For each sample image, the first, second, and third rows show $\mathbf{A}'_i$, $\mathbf{M}_i$, and $\mathbf{A}_i$ for 12 evaluated AUs, respectively.

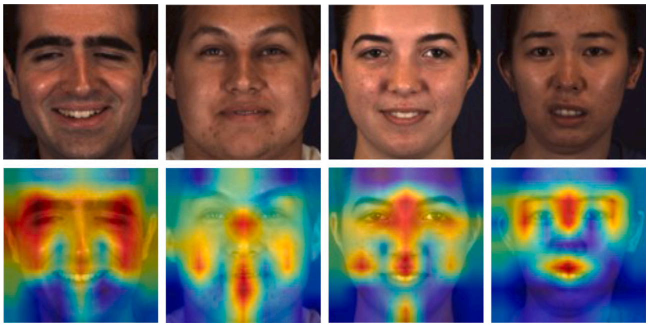


Fig. 5. Visualization of dual-directional attention learned by our GDA on several example images from BP4D [28]. The heatmaps are obtained using the visualization method Grad-CAM [40]. It can be observed that highlighted regions are relevant to such AUs: AUs 6, 10, 12, 14, and 15 in the first column; AUs 1, 2, 6, 7, 10, 12, and 14 in the second column; AUs 7, 10, and 12 in the third column; AUs 4 and 10 in the fourth column.

and GDA. By leveraging a multi-view perspective, ESCDA provides a more comprehensive focus on all categories, thereby alleviating the issue of label imbalance to some extent. For LSC and GDA, notable improvements are observed on low-frequency AUs such as AU 1 and AU 9 compared to their respective variants, indicating their enhanced ability to capture the characteristics of infrequent AUs.

4.4. Visual results

Fig. 4 visualizes pre-constrained self-attention $\mathbf{A}'_i$ averaged over channels, predefined mask $\mathbf{M}_i$, post-constrained self-attention $\mathbf{A}_i$ averaged over channels for each AU learned by our LSC. It can be seen that unconstrained self-attention is dispersed and unable to concentrate on the ROI of each AU. Once combined with the mask, self-attention effectively captures the areas adjacent to the AU. Furthermore, thanks to the learnable parameter ϵ_i, potentially important regions outside the ROI of AU are adaptively captured across different channels, with each channel revealing unique patterns. For instance, regions outside the ROI of AU 7 are learned as higher importance than those of AU 6. Incorporating prior knowledge of AU positions and learnable parameters allows our LSC to not only identify regions strongly related to the AU, but also adaptively detect weakly related areas distant from the ROI.

Fig. 5 visualizes the highlighted areas when our GDA predicts the activation of AUs. We can observe that the model covers the global face with different learned attentions. The model can accurately allocate focused attention to the areas corresponding to the AUs present in the image. For instance, in the first sample, areas corresponding to AU 6 (cheek raiser), AU 10 (upper lip raiser), AU 12 (lip corner puller), AU 14 (dimpler), and AU 15 (lip corner depressor) exhibit significant responses. Furthermore, the attention response areas are continuous, indicating that the model considers both individual AU activation and the spatial relationships among adjacent AUs. This observation aligns with the human understanding that facial expressions are a continuous, dynamic process, not merely a combination of independent AUs.

5. Conclusion

In this paper, we have presented an AU detection framework that assembles a LSC network and a GDA network. By assembling these models, the combination of global and local views allows our framework to identify emotional traits at both comprehensive levels, thereby enhancing its generalizability. We have compared our method against state-of-the-art approaches on the BP4D, DISFA, and GFT datasets, demonstrating significant superiority over prior approaches. Besides, ablation studies further reveal that each key component of our framework plays a role in AU detection. Moreover, visualization results underscore the efficacy of both the LSC network and the GDA network.

Limitation and future work. Our proposed AU detection framework captures the local and global information in static images, but the temporal information contained in the dynamic changes of expressions is also important. In the future work, we will incorporate temporal models such as temporal transformers to capture cross-frame AU relationships, and introduce technologies to optimize attention computation operations, thereby addressing the high computational costs problems.

CRediT authorship contribution statement

Zhiwen Shao: Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization. **Bikuan Chen:** Writing – original draft, Visualization, Methodology, Investigation. **Yong Zhou:** Visualization, Software. **Xuehuai Shi:** Validation, Funding acquisition. **Canlin Li:** Funding acquisition, Data curation. **Lizhuang Ma:** Resources, Formal analysis. **Dit-Yan Yeung:** Writing – review & editing, Validation, Resources, Methodology, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62472424, in part by the China Postdoctoral Science Foundation under Grant 2023M732223, and in part by the Hong Kong Scholars Program under Grant XJ2023037/HKSP23EG01. It was also supported in part by the National Natural Science Foundation of China under Grants 62272461, 62402231, and 72192821, in part by the Opening Fund of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University under Grant VRLAB2024C03, in part by the Research Impact Fund of Hong Kong Government under Grant R6003-21, and in part by the Science and Technology Planning Project of Henan Province under Grant 242102211003.

Data availability

The used BP4D, DISFA, and GFT can be downloaded at http://www.cs.binghamton.edu/~lijun/Research/3DFE/3DFE_Analysis.html, http://mohammadmahoor.com/pages/databases/disfa, and https://osf.io/7wcyz, respectively.

References

[1] G. Xiang, S. Yao, X. Wu, H. Deng, G. Wang, Y. Liu, F. Li, Y. Peng, Driver multi-task emotion recognition network based on multi-modal facial video analysis, Pattern Recognit. (2024) 111241.

[2] J. Mao, R. Xu, X. Yin, Y. Chang, B. Nie, A. Huang, Y. Wang, Poster++: A simpler and stronger facial expression recognition network, Pattern Recognit. (2024) 110951.

[3] P. Ekman, W.V. Friesen, Facial Action Coding System: A Technique for the Measurement of Facial Movement, Consulting Psychologists Press, 1978.

[4] G.M. Jacob, B. Stenger, Facial action unit detection with transformers, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2021, pp. 7680–7689.

[5] W. Li, F. Abtahi, Z. Zhu, L. Yin, EAC-net: Deep nets with enhancing and cropping for facial action unit detection, IEEE Trans. Pattern Anal. Mach. Intell. 40 (11) (2018) 2583–2596.

[6] Z. Shao, Y. Zhou, J. Cai, H. Zhu, R. Yao, Facial action unit detection via adaptive attention and relation, IEEE Trans. Image Process. 32 (2023) 3354–3366.

[7] S. Chen, Y. Liu, X. Gao, Z. Han, Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices, in: Chinese Conference on Biometric Recognition, Springer, 2018, pp. 428–438.

[8] Q. Hou, D. Zhou, J. Feng, Coordinate attention for efficient mobile network design, in: IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp. 13713–13722.

[9] M. Tan, Q.V. Le, MixConv: Mixed depthwise convolutional kernels, in: British Machine Vision Conference, BMVA Press, 2019, p. 74.

[10] Z. Allen-Zhu, Y. Li, Towards understanding ensemble, knowledge distillation and self-distillation in deep learning, in: International Conference on Learning Representations, 2023.

[11] W. Yu, W. Wei, Local and global feature attention fusion network for face recognition, Pattern Recognit. (2024) 111227.

[12] K. Yuan, Z. Yu, X. Liu, W. Xie, H. Yue, J. Yang, Auformer: Vision transformers are parameter-efficient facial action unit detectors, in: European Conference on Computer Vision, Springer, 2025, pp. 427–445.

[13] Z. Li, Z. Zhang, L. Yin, SAT-net: Self-attention and temporal fusion for facial action unit detection, in: International Conference on Pattern Recognition, IEEE, 2021, pp. 5036–5043.

[14] S. Jaiswal, M. Valstar, Deep learning the dynamic appearance and shape of facial action units, in: IEEE Winter Conference on Applications of Computer Vision, IEEE, 2016, pp. 1–8.

[15] K. Zhao, W.-S. Chu, H. Zhang, Deep region and multi-label learning for facial action unit detection, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2016, pp. 3391–3399.

[16] C. Ma, L. Chen, J. Yong, AU R-CNN: Encoding expert prior knowledge into R-CNN for action unit detection, Neurocomputing 355 (2019) 35–47.

[17] Z. Shao, Z. Liu, J. Cai, Y. Wu, L. Ma, Facial action unit detection using attention and relation learning, IEEE Trans. Affect. Comput. 13 (3) (2022) 1274–1289.

[18] X. Liu, K. Yuan, X. Niu, J. Shi, Z. Yu, H. Yue, J. Yang, Multi-scale promoted self-adjusting correlation learning for facial action unit detection, IEEE Trans. Affect. Comput. (2024).

[19] X. Ge, J.M. Jose, S. Xu, X. Liu, H. Han, MGRR-net: Multi-level graph relational reasoning network for facial action unit detection, ACM Trans. Intell. Syst. Technol. 15 (3) (2024) 1–20.

[20] Y. Chen, G. Song, Z. Shao, J. Cai, T.-J. Cham, J. Zheng, Geoconv: Geodesic guided convolution for facial action unit recognition, Pattern Recognit. 122 (2022) 108355.

[21] M. Moghimi, S.J. Belongie, M.J. Saberian, J. Yang, N. Vasconcelos, L.-J. Li, Boosted convolutional neural networks, in: British Machine Vision Conference, BMVA Press, 2016, p. 6.

[22] J. Li, H. Chang, J. Yang, Sparse deep stacking network for image classification, in: AAAI Conference on Artificial Intelligence, 2015, pp. 3804–3810.

[23] W. Jiang, Y. Wu, F. Qiao, L. Meng, Y. Deng, C. Liu, Model level ensemble for facial action unit recognition at the 3rd ABAW challenge, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2022, pp. 2337–2344.

[24] J.-Y. Jeong, Y.-G. Hong, D. Kim, J.-W. Jeong, Y. Jung, S.-H. Kim, Classification of facial expression in-the-wild based on ensemble of multi-head cross attention networks, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2022, pp. 2353–2358.

[25] J. Li, X. Xia, W. Li, H. Li, X. Wang, X. Xiao, R. Wang, M. Zheng, X. Pan, Next-vit: Next generation vision transformer for efficient deployment in realistic industrial scenarios, 2022, arXiv preprint arXiv:2207.05501.

[26] Z. Shao, Z. Liu, J. Cai, L. Ma, JÂa-net: Joint facial action unit detection and face alignment via adaptive attention, Int. J. Comput. Vis. 129 (2) (2021) 321–340.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, Curran Associates, Inc., 2017, pp. 5998–6008.

[28] X. Zhang, L. Yin, J.F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, J.M. Girard, Bp4d-spontaneous: A high-resolution spontaneous 3d dynamic facial expression database, Image Vis. Comput. 32 (10) (2014) 692–706.

[29] S.M. Mavadati, M.H. Mahoor, K. Bartlett, P. Trinh, J.F. Cohn, Disfa: A spontaneous facial action intensity database, IEEE Trans. Affect. Comput. 4 (2) (2013) 151–160.

[30] J.M. Girard, W.-S. Chu, L.A. Jeni, J.F. Cohn, Sayette group formation task (gft) spontaneous facial expression database, in: IEEE International Conference on Automatic Face & Gesture Recognition, IEEE, 2017, pp. 581–588.

[31] Y. Guo, L. Zhang, Y. Hu, X. He, J. Gao, Ms-celeb-1m: A dataset and benchmark for large-scale face recognition, in: European Conference on Computer Vision, Springer, 2016, pp. 87–102.

[32] Y. Chen, D. Chen, T. Wang, Y. Wang, Y. Liang, Causal intervention for subject-deconfounded facial action unit recognition, in: AAAI Conference on Artificial Intelligence, 2022, pp. 374–382.

[33] Y. Chang, S. Wang, Knowledge-driven self-supervised representation learning for facial action unit recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2022, pp. 20417–20426.

[34] X. Li, X. Zhang, T. Wang, L. Yin, Knowledge-spreader: Learning semi-supervised facial action dynamics by consistifying knowledge granularity, in: IEEE International Conference on Computer Vision, IEEE, 2023, pp. 20979–20989.

[35] X. Li, Z. Zhang, X. Zhang, T. Wang, Z. Li, H. Yang, U. Ciftci, Q. Ji, J. Cohn, L. Yin, Disagreement matters: Exploring internal diversification for redundant attention in generic facial action analysis, IEEE Trans. Affect. Comput. 15 (2) (2023) 620–631.

[36] J. Yang, Y. Hristov, J. Shen, Y. Lin, M. Pantic, Toward robust facial action units' detection, Proc. IEEE 111 (10) (2023) 1198–1214, http://dx.doi.org/10.1109/JPROC.2023.3257542.

[37] Y. Li, J. Zeng, S. Shan, X. Chen, Self-supervised representation learning from videos for facial action unit detection, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2019, pp. 10924–10933.

[38] I.O. Ertugrul, J.F. Cohn, L.A. Jeni, Z. Zhang, L. Yin, Q. Ji, Crossing domains for AU coding: Perspectives, approaches, and measures, IEEE Trans. Biom. Behav. Ident. Sci. 2 (2) (2020) 158–171.

[39] Y. Deng, X. Liu, L. Meng, W. Jiang, Y. Dong, C. Liu, Multi-modal information fusion for action unit detection in the wild, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 5855–5862.

[40] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: IEEE International Conference on Computer Vision, 2017, pp. 618–626.