Lock and Check: A Novel Approach for Facial Action Unit Recognition

Journal:	*Transactions on Affective Computing*
Manuscript ID	TAFFC-2025-03-0229
Manuscript Type:	Regular
Opposed Editors:	
Keywords:	I.4 Image Processing and Computer Vision < I Computing Methodologies, O Affective Computing, O.1 Affect sensing and analysis < O Affective Computing

Lock and Check: A Novel Approach for Facial Action Unit Recognition

Zihao Huang, Jian Gao, Wentian Cai, Yandan Chen, Xiping Hu, Ping Gao, and Ying Gao, *Member, IEEE*

***Abstract*—Facial Action Unit (AU) recognition involves identifying subtle muscle movements corresponding to different AUs. Recent approaches have focused on localizing AUs using predefined Regions of Interest (RoIs) or learnable modules. However, these methods either overly depend on the precision of predefined RoIs or inaccurately localize background regions instead of the actual AU positions. To address this challenge, we propose a novel method, Lock and Check (LAC), which automatically localizes each AU without relying on predefined RoIs or introducing learnable modules during the inference phase. Specifically, our approach decomposes the task into two subtasks: AU localization and AU state verification. We first align the direction between spatial features and the corresponding AU class weights to guide the model in localizing AUs. Next, we incorporate spatial and temporal aspects for precise AU state detection. From the perspective of spatial information learning, we propose a confidence-based AU relationship mining module that directs the model to focus on uncertain AUs. From the aspect of temporal information learning, we introduce a temporal sampling strategy that implicitly captures time-dependent features. Experimental results on the BP4D and DISFA datasets demonstrate the effectiveness of our method, showing that it outperforms existing approaches and achieves state-of-the-art performance in AU recognition.**

***Index Terms*—Facial action unit recognition, action unit localization, relation modeling, spatial and temporal learning.**

I. Introduction

FACIAL action units (AUs), introduced by the Facial Action Coding System (FACS) [1], represent the movements of specific facial muscles. Nearly all facial behaviors can be described by various combinations of AUs. Facial AU recognition has garnered significant attention due to its potential applications in healthcare, gaming, and beyond. This task is inherently a multi-label classification problem, as one or more AUs can be activated simultaneously. It has been proven that local regions are crucial for AU recognition, as AUs activate in specific facial areas [2].

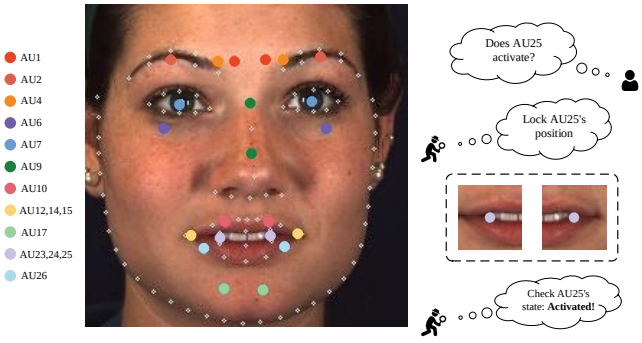

Fig. 1. Visualization of AU-related center points and the pipeline of LAC.

Recent methods [3]–[6] have attempted to localize AUs using predefined RoIs or learnable modules. Some methods [3]–[5] predefine AU heatmaps based on preprocessed facial landmarks for AU localization. Moreover, [6], [7] divide the image into several areas based on hypotheses or prior knowledge. Meanwhile, other methods employ learnable modules, such as attention mechanisms [8] and shallow RoI detection networks [9], to localize AUs. However, these approaches either overly rely on the precision of predicted RoIs or fail to localize AUs accurately in the presence of noise. Furthermore, simply localizing AUs is insufficient to determine whether they are activated. Previous studies have addressed this issue by incorporating temporal information [6], [10] or modeling AU relationships [9], [11]–[13].

Integrating temporal information [14]–[16] has proven to be effective for facial AU recognition. Although facial AUs activate at the same positions and produce similar patterns on the face, detecting AUs with subtle motions remains challenging. Therefore, incorporating temporal features is essential. A natural approach for leveraging temporal information is to utilize temporal models such as Long Short-Term Memory (LSTM) [17], [18] Networks, temporal Graph Convolutional Networks (GCNs) [14], [19], temporal difference networks [6], [10], and 3D Convolutional Neural Networks (CNNs) [20], [21] to capture AU motions. However, none of these methods have explored the impact of sequence sampling strategies, which we found is crucial for AU recognition, even for still image classification.

In addition to temporal information coordination, a significant portion of studies [22]–[26] has focused on AU relationship modeling. These studies [6] assume that aggregating information from related AUs can help the model identify AUs

This work is supported by Guangzhou's Key Areas R&D Special Topic (Industrial Chain) Project (No. 2024B01W0029).

Zihao Huang, Jian Gao, Wentian Cai, Yandan Chen, and Ying Gao are with the School of Computer Science and Engineering, South China University of Technology, Guangzhou, 510006, P. R. China and Guangdong Provincial Key Laboratory of Artificial Intelligence in Medical Image Analysis and Application (e-mail: {cshzih, csgaojian, cscaiwentian, csydchen}@mail.scut.edu.cn, gaoying@scut.edu.cn).

Ping Gao is with the Department of Geriatric Respiratory Medicine, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, and the Guangdong Provincial Geriatrics Institute, Guangzhou, 510080, Guangdong, China (e-mail: gaoping8599@gdph.org.cn).

Xiping Hu is with the School of Medical Technology, Beijing Institute of Technology, Beijing 100081, China, and also with the AI Research Institute, Shenzhen MSU-BIT University, Shenzhen 518172, China (e-mail: huxp@bit.edu.cn)

Corresponding author: Ying Gao (e-mail: gaoying@scut.edu.cn), and Ping Gao (e-mail: gaoping8599@gdph.org.cn).

with low confidence. Some approaches [27], [28] build AU graphs based on existing AU-related knowledge, such as FACS definitions and anatomical information. While others [9], [13], [15] use metrics like cosine similarity or Euclidean distance to measure the proximity between AU features, and then apply a top-K or threshold strategy to connect highly relevant AU pairs. Both types of studies have attempted to use GCNs [29] or their variants to propagate information between AU pairs. However, a challenge remains: the confidence of classes with high initial confidence (i.e., those with probabilities close to 0 or 1) may be reduced after GCN operations.

In this paper, we propose a novel approach, Lock and Check (LAC), which (i) explicitly models AU positions without using learnable modules, and (ii) detects AU states (i.e., whether activated or not) through both spatial and temporal modeling. Our LAC framework consists of three key components. First, we decompose AU recognition into two subtasks: AU localization and AU state determination. We align the direction between spatial features and the corresponding AU class weights to guide the model's focus on the region where the AU is most likely to appear. Second, we introduce a temporal sampling strategy that accounts for subject and class diversity, implicitly capturing time-dependent features to enhance AU motion detection. Third, the confidence-based AU relationship mining module resolves the states of uncertain AUs with low confidence (i.e., those with probabilities around 0.5). Finally, we incorporate LSTM networks to model temporal information explicitly.

To summarize, the main contributions of our work are as follows:

- We align the direction between AU classes' weights and the corresponding spatial features to direct the model to focus on regions where AUs are likely to activate.
- We propose a temporal sampling strategy that implicitly captures time-dependent features, enhancing AU recognition performance, even for still image classification.
- We propose a confidence-based AU relationship mining module that helps resolve AUs with low confidence.
- Experiments conducted on the widely used datasets: BP4D [30] and DISFA [31] show that LAC can outperform existing methods and achieve state-of-the-art performance in recognizing AU.

II. Related Works

A. ROI-Based AU Recognition

Different AUs are associated with various facial regions. In contrast to global facial regions, local facial areas can more precisely describe AUs and reduce background noise. In earlier work, [2], [7], [22], [32], [33] cropped input images into multiple grids and extracted features from each Region of Interest (RoI). To obtain more precise RoIs, [34]–[37] defined RoIs based on facial landmarks and anatomical knowledge. Rather than using predefined RoI boxes, [3], [4], [38] transformed facial landmarks into heatmaps for RoI extraction, refining these heatmaps to capture precise local features. Instead of preset heatmaps, [9] attempted to obtain AU-specific regions through importance maps generated from global features. Similarly, [39] proposed a RoI attention module to extract areas containing active AUs. [8] estimated RoIs based on a cross-modality attention mechanism. Moreover, [23] introduced a channel-wise and spatial attention learning module to capture activated regions on the face. Additionally, [40] employed generative networks to learn facial landmarks. In contrast to the above methods, [5] randomly masked RoIs corresponding to specific AUs and attempted to recover the masked regions to explore the relationships between AU pairs. In this paper, we design loss functions to align the direction between AU class weights and the corresponding spatial features, guiding the model in localizing AUs without relying on any learnable modules. Moreover, our method can focus on the specific regions where AUs activate during the inference stage, without predefined RoIs.

B. Temporal Information-Based AU Recognition

Temporal information plays a crucial role in AU recognition, particularly for AUs with subtle movements. In pioneering work, [41], [42] employed machine learning techniques such as Support Vector Machine (SVM) and Dynamic Bayesian Network (DBN) to model the temporal changes of activated AUs. With the advancement of deep learning, [17], [18] utilized LSTM networks to capture temporal information following spatial feature extraction. [43] further introduced Conv-LSTM to concurrently extract spatial and temporal features. Beyond LSTM, [20], [21] directly applied 3DCNNs to process AU image sequences, eliminating the need for complex architectural designs. To reduce the framework's parameters, [44] replaced 3DCNNs with 2+1D CNNs to model spatial-temporal information. Additionally, [6], [10] simplified the model architecture by employing temporal difference networks, which directly subtract features between consecutive frames. Instead of CNNs and LSTMs, Transformers [45] have emerged as an advanced tool for processing sequence data. [19] used a temporal transformer to model temporal information. [19] further proposed a spatial-temporal transformer to capture AU motions by combining both spatial and temporal features. In addition to these methods, dynamic facial shapes [46], [47] and optical flow [5], [48], [49] were also estimated to represent temporal changes in facial movements. Moreover, head motion is often unavoidable, which can degrade AU recognition performance. To address this issue, [50] proposed a twin-cycle autoencoder to disentangle AU and head motions. However, none of the methods mentioned above have investigated the impact of temporal sampling strategies, which is one of the key contributions of this paper.

C. Relaion-Based AU Recognition

Injecting AU relationships into AU recognition can significantly enhance performance. There are several ways to constructing AU relationship graphs. One type of method derives AU relationship graphs from prior knowledge. For instance, [25] predefined an AU relationship graph based on rules established in the FACS to facilitate learning. Similarly,

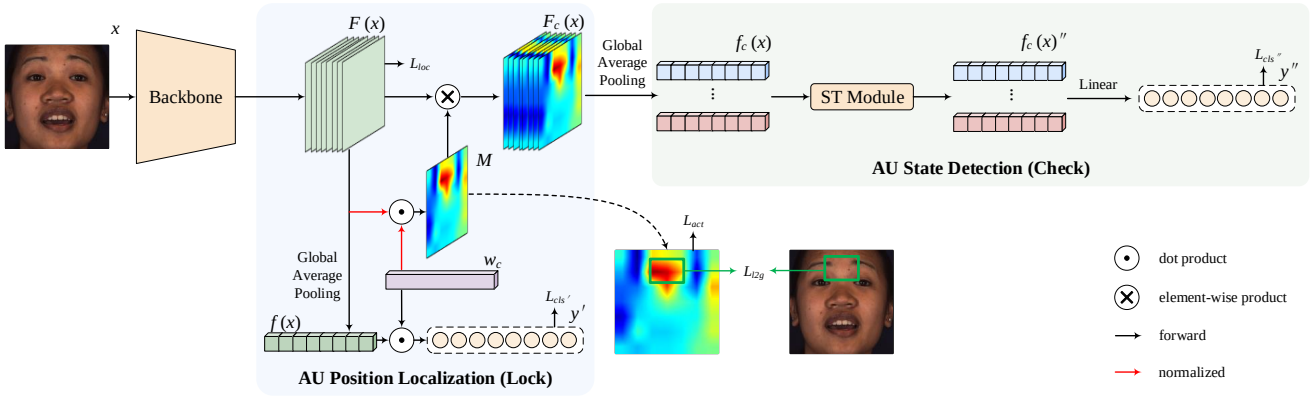

Fig. 2. **Illustration of LAC framework**: It mainly consists of two steps: AU position localization and AU state detection, which correspond to the Lock and Check steps. During the Lock step, our model learns to localize AUs and generates AU-specific heatmaps for local feature extraction. Using these AU features, our model detects whether AUs activate in the Check step with the proposed Spatial-Temporal (ST) Module.

[27], [51] constructed graphs based on the relationships between emotions and AU cooccurrences recorded in EMFACS [52]. Furthermore, [22] predefined graphs based on AUs' correlations to explore all potential AUs. However, static AU relationship graphs may introduce bias into the model's training process by evoking unactivated AUs based on their connections to activated AUs. To address this issue, another category of methods has attempted to build AU relationship graphs in a learnable manner. [9], [13], [23], [24] employed multiple AU descriptors to describe various AU features, then designed modules such as Graph Attention Networks (GATs) [53] to learn AU relationships. [15] further introduced uncertain graph convolution, which simultaneously estimates AUs' dependencies and the uncertainties of the predictions. Moreover, [28] updated their relationship graph based on the initial predicted probabilities of AUs. Additionally, [8] explored AU relationships using an attention module to process embeddings of AU descriptive text. To learn subject-invariant representations, [54] maintained consistency in unary, binary, and multivariate AU relationships across diverse subjects. However, most relation-based methods directly employ GCNs or their variants to process AU features, which may weaken the confidence of AUs that have high confidence (i.e., those with predicted probabilities close to 0 or 1). In this paper, we propose a confidence-based AU relationship mining module to address this problem.

The most related work is [4], where facial priors constrain the RoI learning process during training, and AUs are localized without priors during the inference stage. However, our method differs in the following ways: (1) [4] proposed additional modules for facial landmark detection to generate RoI heatmaps based on these landmarks, while our approach localizes AUs using well-defined loss, without the need for extra modules. (2) Unlike [4], which combines global and local features and designs subsidiary task, our work only focuses on local feature modeling. (3) Our work incorporates temporal and AU relationship modeling, aspects not considered by [4].

III. Method

The overall flow of the proposed **LAC** method for facial AU recognition is shown in Figure 2. It consists of two main steps: AU position localization and AU state detection. The input image $x \in \mathbb{R}^{H\times W\times 3}$ is first encoded into an image feature $F(x) \in \mathbb{R}^{N\times H'\times W'}$ by the backbone. Next, $F(x)$ is passed to two branches. One branch estimates the AU-specific heatmaps $M \in \mathbb{R}^{N_c\times H'\times W'}$ for N_c classes by performing a dot product between $F(x)$ and the weights $w_c \in \mathbb{R}^{N_c\times N}$ of the linear layer. The other branch calculates AU-specific features $F_c(x) \in \mathbb{R}^{N_c\times N\times H'\times W'}$ by performing an element-wise product between M and $F(x)$. These features, $F_c(x)$, are then converted into AU-specific vectors $f_c(x) \in \mathbb{R}^{N_c\times N}$. The proposed Spatial-Temporal (ST) Module takes $f_c(x)$ as input and returns $f_c(x)'' \in \mathbb{R}^{N_c\times N}$, which contains both spatial and temporal information. Finally, $f_c(x)''$ is used to estimate the probabilities for the AU classes.

A. AU Position Localization

Facial AU recognition aims to maximize the probabilities of activated AUs and suppress the probabilities of unactivated AUs. Given a typical network consisting of convolutional layers, a Global Average Pooling (GAP) layer, and a linear layer, the prediction $y' \in \mathbb{R}^{N_c}$ can be formulated as

$$f(x) = \frac{\sum_i^{H'}\sum_j^{W'} F(x)_{i,j}}{H'\times W'} \tag{1}$$

$$\begin{aligned} y' &= w_c f(x) \\ &= \frac{1}{H'\times W'}\sum_i^{H'}\sum_j^{W'} w_c F(x)_{i,j} \end{aligned} \tag{2}$$

TABLE I
CENTER COORDINATES OF EACH AU CLASS

Regions	AU Classes	Center Coordinates			
		x_{c1}	y_{c1}	x_{c2}	y_{c2}
Upper Face	1	x_{37}	y_{37}	x_{42}	y_{42}
	2	x_{40}	y_{40}	x_{48}	y_{48}
	4	$\overline{x_{37,51}}$	y_{37}	$\overline{x_{42,51}}$	y_{42}
	7	$\overline{x_{60:68}}$	$\overline{y_{60:68}}$	$\overline{x_{68:76}}$	$\overline{y_{68:76}}$
Middle Face	6	x_{67}	y_{52}	x_{73}	y_{52}
	9	x_{51}	y_{51}	x_{53}	y_{53}
Lower Face	10	$\overline{x_{77,78}}$	$\overline{y_{77,78}}$	$\overline{x_{80,81}}$	$\overline{y_{80,81}}$
	12	x_{88}	y_{88}	x_{92}	y_{92}
	14	x_{88}	y_{88}	x_{92}	y_{92}
	15	x_{88}	y_{88}	x_{92}	y_{92}
	17	$\overline{x_{15,86}}$	$\overline{y_{15,86}}$	$\overline{x_{17,84}}$	$\overline{y_{17,84}}$
	23	$\overline{x_{89,95}}$	$\overline{y_{89,95}}$	$\overline{x_{91,93}}$	$\overline{y_{91,93}}$
	24	$\overline{x_{89,95}}$	$\overline{y_{89,95}}$	$\overline{x_{91,93}}$	$\overline{y_{91,93}}$
	25	$\overline{x_{89,95}}$	$\overline{y_{89,95}}$	$\overline{x_{91,93}}$	$\overline{y_{91,93}}$
	26	x_{83}	y_{83}	x_{87}	y_{87}

$*$ *x_i and y_i denote the coordinates of the ith facial landmark, $\overline{x_{i,j}}$ represents $(x_i + x_j)/2$, and $\overline{x_{i:j}}$ is the average value from x_i to x_{j-1}. Similarly, y is defined in the same way.*

where $f(x)$ denotes the AU vectors after the GAP operation. We can further decompose y' as

$$y' = \frac{1}{H' \times W'} \sum_{i}^{H'} \sum_{j}^{W'} |w_c| \, |F(x)_{i,j}| \cos(w_c, F(x)_{i,j}) \quad (3)$$

where $|w_c|$ is the magnitude of w_c, $|F(x)_{i,j}|$ denotes the norm of $F(x)_{i,j}$, and $\cos(w_c, F(x)_{i,j})$ represents the cosine similarity between w_c and $F(x)_{i,j}$. It can be seen that y' is determined by $|w_c|$, $|F(x)_{i,j}|$, and $\cos(w_c, F(x)_{i,j})$. Since $|w_c|$ is fixed for each sample, y' is strongly related to $|F(x)_{i,j}|$ and $\cos(w_c, F(x)_{i,j})$. Based on this observation, we aim to highlight the regions where AUs are likely to activate by maximizing $|F(x)_{i,j}|$ and localize activated AUs by maximizing $\cos(w_c, F(x)_{i,j})$ in areas where AUs are activated while minimizing $\cos(w_c, F(x)_{i,j})$ in regions where AUs do not activate. Therefore, the direction between the activated AU classes' weights and the corresponding spatial features will be aligned, directing the model to focus on the areas where AUs are likely to appear.

Specifically, we first predefined RoIs $R_c(x) \in \mathbb{R}^{N_c \times 2 \times 4}$ based on landmarks predicted by STAR Loss [55]. Each AU class consists of two RoIs, determined by two center points, as detailed in Table I. We then proceed to calculate $|F(x)| \in \mathbb{R}^{H' \times W'}$ and $\cos(w_c, F(x)) \in \mathbb{R}^{N_c \times H' \times W'}$. To perform AU localization, we apply RoIAlign [56] to $|F(x)|$ and define the loss as

$$L_{loc} = - \sum_{r \in R_c(x)} \sum_{pt \in r} |F(x)_{pt}| \quad (4)$$

where pt represents the point in r. With this function, we highlight the regions where AUs are likely to appear and guide the model to focus on these key areas. After localizing the AUs, we perform RoIAlign on $\cos(w_c, F(x))$ and define the loss as

$$L_{act} = - \sum_{i \in C_{pos}} \sum_{r \in R_i} \sum_{pt \in r} M_{i,pt} + \sum_{j \in C_{neg}} \sum_{r \in R_j} \sum_{pt \in r} M_{j,pt} \quad (5)$$

where $M = \cos(w_c, F(x))$, C_{pos} refers to the indices of activated AU classes, C_{neg} represents the indices of unactivated AUs, R_i denotes the RoIs for the i-th classes, and $M_{i,pt}$ is the value at the pt position in the similarity map of the i-th class. In this loss function, we enhance the similarity between w_c and $F(x)$ for the activated AUs, while suppressing the similarity for unactivated AUs. This enables the model to learn how to distinguish whether the AU is activated or not. Moreover, we suppose that the global similarity score should align with the local score, and we propose a consistency loss as

$$L_{l2g} = -\frac{1}{2N_c} \sum_{i}^{N_c} \sum_{r \in R_i} (\text{GAP}(M_{i,r}) - \text{GAP}(M_i))^2 \quad (6)$$

where $M_{i,r}$ denotes the values of the area r in the i-th class's similarity map M_i and GAP is the global average pooling operation. After performing the aforementioned operations, our model gains a preliminary understanding of AUs. We can then extract AU-specific features $F_c(x)$ by applying the similarity maps M to $F(x)$.

Algorithm 1 Temporal Sampling Strategy

Input: video sequences $S = \{s_1, s_2, \ldots, s_{N_s}\}$ collected from N_s subjects, each sequence $s_i = \{f_1, f_2, \ldots, f_{N_{F_i}}\}$ contains N_{F_i} frames
Parameter: sliding window size t, number of subsequences N_{sep}
Output: video clips $VC = \{vc_1, vc_2, \ldots, vc_{N_c}\}$

```
Let VC = {}.
for i = 1 to N_s do
    N_{SF_i} = ⌊N_{F_i} / N_sep⌋
    for j = 1 to N_sep do
        vc_j = {}
        index_start = RandomChoice(1, N_{SF_i} − t)
        for k = index_start to index_start + t do
            vc_j = vc_j ∪ f_k
        end for
        VC = VC ∪ vc_j
    end for
end for
return VC
```

B. AU State Detection

Temporal Sampling Strategy. Before performing temporal learning, we first need to sample image sequences. An intuitive idea is to directly use a sliding window with a fixed size and stride to collect sequence data. However, this sampling method is inefficient and suffers from data redundancy challenges. Specifically, the collected subsequences of the same object may exhibit high similarity, and many subsequences may lack any activated AUs in the AU recognition task. We hypothesize that two factors are crucial for efficient data sampling: the

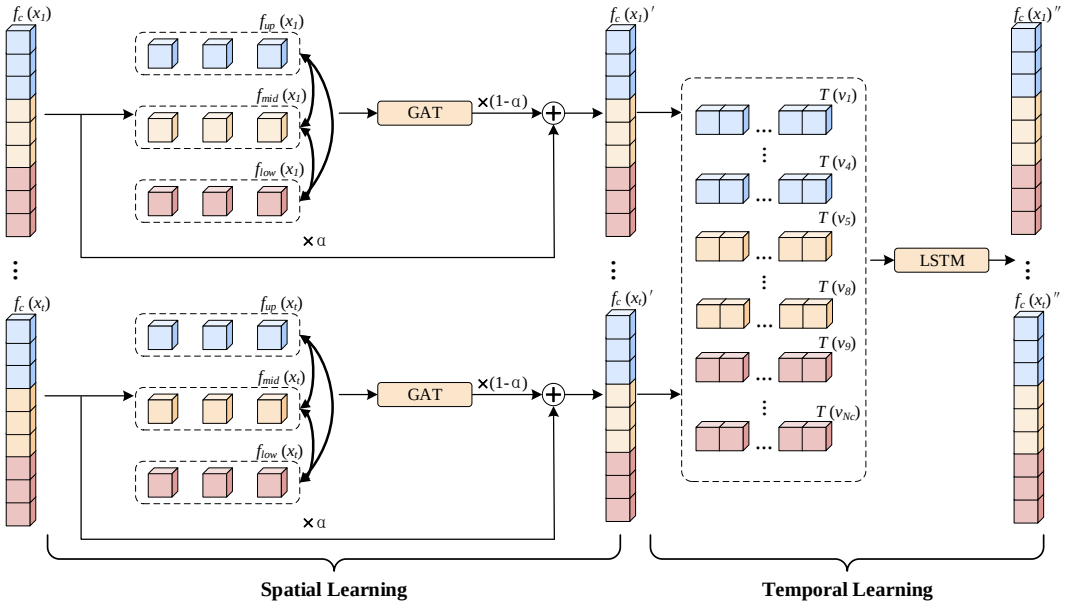

Fig. 3. **Overview of ST Module**: It primarily consists of two steps: spatial learning, which models the AU relationships, and temporal learning, which further extracts temporal features.

sliding window size t and data diversity. A larger t allows the model to capture facial motion better, while a higher degree of data diversity can help prevent overfitting, improving both the model's generalization and performance [57].

To increase data diversity, we first sample each subject equally to ensure subject diversity. We then randomly select the index of the first frame of each clip to introduce clip diversity. Additionally, we partition each subject's sequence into N_{sep} subsequences to enhance sample diversity. The larger the value of N_{sep}, the more diverse the data becomes. Further details of the sampling strategy are outlined in Algorithm 1.

Moreover, we suggest that temporally sampled images can also enhance the performance of still image-based AU classification. A sequence of continuous frames helps reduce background motion noise, allowing the model to implicitly learn time-dependent and AU motion-related features.

Spatial-Temporal (ST) Module. After constructing AU-specific vectors $f_c(x)$ using the GAP operation, we perform spatial-temporal learning to enhance AU state detection, as illustrated in Figure 3. In the spatial learning phase, we concentrate on modeling AU relationships. Motivated by [6], we seperate AUs into three regions: upper face $f_{up}(x) \in \mathbb{R}^{N_{up} \times N}$, middle face $f_{mid}(x) \in \mathbb{R}^{N_{mid} \times N}$ and lower face $f_{low}(x) \in \mathbb{R}^{N_{low} \times N}$, as shown in Table I. We then perform cross-regional AU information transfer with GAT as

$$e_{ij} = \text{LeakyReLU}(a[v_i W || v_j W]) \tag{7}$$

$$\gamma_{ij} = \frac{\exp(e_{ij})}{\sum_k^{N_v} \exp(e_{ik})} \tag{8}$$

$$v_i' = \theta(\sum_{j \in Neb_i} \gamma_{ij} v_j W) \tag{9}$$

where $e_{i,j} \in \mathbb{R}^1$ represents the weight of the edge connecting node i and node j, a denotes the attention mechanism weight, $\gamma_{ij} \in \mathbb{R}^1$ refers to the attention aggregation weight for $v_j \in \mathbb{R}^{1 \times N}$, $||$ denotes the feature concatenation operation, $W \in \mathbb{R}^{N \times N}$ is the weight for node feature update, Neb_i is the neighbour nodes of node i and θ is the activation function.

However, directly applying GAT to all AUs may introduce a problem: the confidence of high-confidence AUs (i.e., those with predicted probabilities close to 0 or 1) may be weakened. To address this, we propose a confidence-based AU relationship mining module that directs the model to focus on uncertain AUs during AU relationship learning. Specifically, we first calculate the confidence score α based on the predicted score y' in the AU position localization step as

$$\alpha = \text{Abs}(y' - 0.5)/0.5 \tag{10}$$

where $\text{Abs}(\cdot)$ refers to the absolute value operation. We then update $f_c(x)$ while preserving the confidence of the high-confidence AUs as

$$f_c(x)' = f_c(x) \times \alpha + \text{GAT}(f_c(x)) \times (1 - \alpha) \tag{11}$$

After spatial learning, we collect processed AU-specific feature sequences $\{f_c(x_1)', f_c(x_2)', \ldots, f_c(x_t)'\} \in \mathbb{R}^{t \times N_c \times N}$ and create temporal sequences $\{T(v_1), T(v_2), \ldots, T(v_{N_c})\} \in \mathbb{R}^{N_c \times t \times N}$. These sequences are then passed to the LSTM to extract temporal features, resulting in enhanced AU-specific features $\{f_c(x_1)'', f_c(x_2)'', \ldots, f_c(x_t)''\} \in \mathbb{R}^{t \times N_c \times N}$. Finally, these features are forwarded to a linear layer to obtain the final predictions $\{y_1'', y_2'', \ldots, y_t''\} \in \mathbb{R}^{t \times N_c}$.

C. Loss Function

Previous work [39] has shown that DISFA and BP4D are unbalanced datasets. To address this issue, [6], [13] leverage

TABLE II
F1 SCORE OF FACIAL AU RECOGNITION ON THE BP4D DATASET

Method	AU												Avg.
	1	2	4	6	7	10	12	14	15	17	23	24	
RoI-Based Methods													
DRML [2]	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
$JÂA$-Net [4]	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
AAR [38]	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
GLEE-Net [35]	60.6	44.4	61.0	80.6	78.7	**85.4**	88.1	64.9	53.7	65.1	47.7	58.5	[65.7]
MPSC [37]	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
Temporal Information-Based Methods													
BGAD [47]	57.4	**52.6**	**64.6**	79.3	**81.5**	82.7	85.6	67.9	47.3	58.0	47.0	44.9	64.1
AUNet [10]	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
MDHR [6]	[58.3]	50.9	58.9	78.4	[80.3]	84.9	88.2	69.5	[56.0]	**65.5**	49.5	**59.3**	66.6
Relation-Based Methods													
SRERL [22]	46.9	45.3	55.6	77.1	78.4	83.5	87.6	63.9	52.2	63.9	47.1	53.3	62.9
MEGraph [13]	52.7	44.3	60.9	[79.9]	80.1	85.3	**89.2**	[69.4]	55.4	64.4	49.8	55.1	65.5
SupHCL [54]	52.8	45.7	61.6	79.5	79.3	84.7	86.9	67.6	51.4	62.5	48.6	52.3	64.4
KDSSRL [25]	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
Other Methods													
SMA [58]	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
ETD [59]	54.7	[50.8]	57.1	78.8	79.6	84.6	88.0	67.0	54.9	62.9	48.6	54.5	65.1
LAC (ours)	**61.5**	49.6	[61.8]	79.2	79.7	[85.2]	88.0	**70.1**	**58.4**	[64.8]	[51.7]	[57]	**67.3**

** The best, second best, and third best results of each column are shown with **bold font**, underline, [bracket] respectively.*

the Weighted Asymmetric (WA) loss, formulated as

$$L_{\text{WA}_i} = w_i[y_i \log(p_i) + (1 - y_i)p_i \log(1 - p_i)] \tag{12}$$

where w_i refers to the weight for the i-th class, estimated based on the entire training set, y_i represents the ground truth of the i-th class, and p_i is the predicted probability of the i-th class. However, this approach is unsuitable for our method, as temporal sampling may modify the data distribution in each epoch. Therefore, we propose a Dynamic Weighted Asymmetric (DWA) Loss, defined as

$$w_{\text{diff}_{i,e+1}} = \beta_1 w_{\text{diff}_{i,e+1}} + (1 - \beta_1)\frac{1}{\text{Recall}_{i,e}} \tag{13}$$

$$w_{\text{dist}_{i,e+1}} = \beta_2 w_{\text{dist}_{i,e+1}} + (1 - \beta_2)\frac{1}{\text{ClsDist}_{i,e}} \tag{14}$$

$$w_{i,e} = \delta w_{\text{diff}_{i,e+1}} + (1 - \delta) w_{\text{dist}_{i,e+1}} \tag{15}$$

$$L_{\text{DWA}_{i,e}} = w_{i,e}[y_{i,e} \log(p_{i,e}) + (1 - y_{i,e})p_{i,e} \log(1 - p_{i,e})] \tag{16}$$

where $\text{diff}_{i,e}$ refers to the i-th class's recognition difficulty weight in epoch e, $\text{dist}_{i,e}$ denotes the i-th class's distribution weight in epoch e, $\text{Recall}_{i,e} = \frac{TP_{i,e}}{TP_{i,e}+FN_{i,e}}$ represents the i-th class's recall rate estimated in epoch e, $\text{ClsDist}_{i,e} = \frac{N_{i,e}}{N_{C,e}}$ is the i-th class's distribution estimated in epoch e, and $\beta 1, \beta 2, \delta$ are predefined coefficients. Compared to the WA loss, our loss function introduces a dynamic weight updated in each epoch. The DWA loss considers data distribution and prediction difficulty (i.e., recall rate), leading to improved performance.

The final loss is given by

$$\begin{aligned} L =& \eta_1 L_{loc} + \eta_2 L_{act} + \eta_3 L_{l2g} \\ &+ \eta_4 L_{\text{DWA}}(y', y) + \eta_5 L_{\text{DWA}}(y'', y) \end{aligned} \tag{17}$$

where $\eta_1, \eta_2, \eta_3, \eta_4, \eta_5$ are predefined coefficients.By optimizing L, our model can automatically localize AUs, check AU states, and achieve promising results in facial AU recognition.

IV. EXPERIMENTS

A. Datasets and Evaluation Methods

We evaluate our approach on the widely-used AU recognition datasets: DISFA and BP4D. DISFA contains 130,815 frames collected from 27 subjects (12 females and 15 males). Each frame is annotated with six-point scale intensity labels for 12 AUs. BP4D consists of 146,847 images recorded from 41 subjects (23 females and 18 males). Each subject is asked to participate in 8 tasks to induce different expressions. Similar to [2], we adopt three-fold subject-independent cross-validation for evaluation. To compare performance with the state-of-the-art methods, we use the common metric: frame-based F1 score.

B. Implementation Details

Facial landmarks detection and face alignment are performed for each frame. All face images are resized to 224 $\times$ 224. Resnet 50 [60] pre-trained on ImageNet [61], is used as the backbone. During training, input images are processed with data augmentation, including random horizontal flipping and color jitter. Additionally, several RoIs are generated as described in Table I during the training phase, while the model can perform inference without RoIs. We use the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay of $5e^{-4}$ to optimize other learnable parameters. The model is trained for 20 epochs with an initial learning rate of $1e^{-4}$ for still image-based AU recognition. For video clip-based AU recognition, the model is trained for 200 epochs with the same initial learning rate. The batch size is set as 64 for image-based classification and 8 for clip-based

TABLE III
F1 SCORE OF FACIAL AU RECOGNITION ON THE DISFA DATASET

Method	AU								Avg.
	1	2	4	6	9	12	25	26	
RoI-Based Methods									
DRML [2]	17.3	17.7	37.4	29.0	10.7	37.7	38.5	20.1	26.7
$J\hat{A}A$-Net [4]	[62.4]	[60.7]	67.1	41.1	45.1	73.5	90.9	67.4	63.5
AAR [38]	[62.4]	53.6	71.5	39.0	48.8	76.1	91.3	**70.6**	64.2
GLEE-Net [35]	61.9	54.0	**75.8**	45.9	55.7	[77.6]	92.9	60.0	65.5
MPSC [37]	62.0	**65.7**	74.5	53.2	43.1	76.9	**95.6**	53.1	65.5
Temporal Information-Based Methods									
BGAD [47]	41.5	44.9	60.3	51.5	50.3	70.4	91.3	55.3	58.2
AUNet [10]	60.3	59.1	69.8	48.4	53.0	**79.7**	93.5	[64.7]	[66.1]
MDHR [6]	**65.4**	60.2	75.2	50.2	[52.4]	74.3	93.7	58.2	66.2
Relation-Based Methods									
SRERL [22]	45.7	47.8	59.6	47.1	45.6	73.5	84.3	43.6	55.9
MEGraph [13]	54.6	47.1	72.9	[54.0]	55.7	76.7	91.1	53.0	63.1
SupHCL [54]	52.5	58.8	70.0	53.5	51.4	73.1	**95.6**	58.0	64.1
KDSSRL [25]	60.4	59.2	67.5	52.7	51.5	76.1	91.3	57.7	64.5
Other Methods									
SMA [58]	53.4	54.2	64.0	**57.0**	47.0	76.6	92.0	55.2	62.4
ETD [59]	62.6	54.7	70.8	46.3	51.7	76.3	94.4	59.8	64.6
LAC (ours)	59.3	62.1	[73.7]	55.3	**56.3**	79.1	[93.9]	62.4	**67.8**

** The best, second best, and third best results of each column are shown with **bold font**, underline, [bracket] respectively.*

TABLE IV
AVERAGE F1 SCORE RESULTS WITH DIFFERENT SETTINGS ON BP4D

Backbone	TS	L_{loc}	L_{act}	L_{l2g}	ST	L_{DWA}	F1
✓							62.6
✓	✓						65.1
✓	✓				✓		65.7
✓	✓	✓	✓				65.8
✓	✓	✓	✓	✓			66.1
✓	✓	✓	✓	✓	✓		66.5
✓	✓	✓	✓	✓	✓	✓	**67.3**

** TS refers to temporal sampling strategy, ST represents ST module.*

TABLE V
ABLATION STUDIES OF SLIDING WINDOW SIZE t ON BP4D

t	2	4	8	16	32	64
F1 Score	62.6	62.7	64.0	**65.1**	64.8	64.7

TABLE VI
ABLATION STUDIES OF NUMBER OF SUBSEQUENCES N_{sep} ON DISFA

N_{sep}	1	2	4	8	12	16
F1 Score	54.5	59.6	58.8	61.4	**65.1**	63.1

classification. The values of $w_{\text{diff}_{i,0}}$ and $w_{\text{dist}_{i,0}}$ are initialized with the i-th class's distribution in the training set. The hyperparameters $\{\beta_1, \beta_2, \delta, \eta_1, \eta_2, \eta_3, \eta_4, \eta_5\}$ are empirically set to $\{0.9, 0.9, 0.5, 0.5, 0.5, 0.01, 0.5, 1\}$.

The framework is implemented in Pytorch[1] and NVIDIA GeForce GTX 3090 GPUs are used.

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

In this section, we compare our method with alternative approaches, including RoI-based methods [2], [4], [35], [37], [38], temporal information-based methods [6], [10], [47], relation-based methods [13], [22], [25], [54], and other advanced methods [58], [59]. Tables II and III present the F1 score results on the BP4D and DISFA datasets, respectively. It is evident that our method outperforms all the listed methods on both datasets, achieving new state-of-the-art (SOTA) performance with 67.3% and 67.8% on BP4D and DISFA, respectively. Specifically, our method surpasses the best RoI-based method [35] by 1.6% and 2.3% on BP4D and DISFA, respectively. Moreover, our method outperforms the best relation-based method [13], [25] by 1.8% and 3.3% on BP4D and DISFA, respectively. Additionally, our method achieves improvements of 0.7% and 1.6% over the current SOTA method, MDHR [6], on BP4D and DISFA, respectively. Our method also ranks among the top three for 8 out of 12 AUs (i.e., AU1, AU4, AU10, AU14, AU15, AU17, AU23, AU24) on BP4D and 6 out of 8 AUs (i.e., AU2, AU4, AU6, AU9, AU12, AU25) on DISFA.

Both tables demonstrate the feasibility and robustness of LAC in AU recognition, as it achieves SOTA performance on the BP4D and DISFA datasets. Furthermore, LAC can automatically infer during the inference phase, mitigating the issues caused by facial landmark errors and resulting in performance gains compared to predefined RoI-based methods [37]. Meanwhile, LAC introduces precise RoI information during training to guide the model in better localizing AUs, a capability often neglected by learnable RoI-based methods [38], leading to improved performance. Additionally, the

[1]https://www.pytorch.org

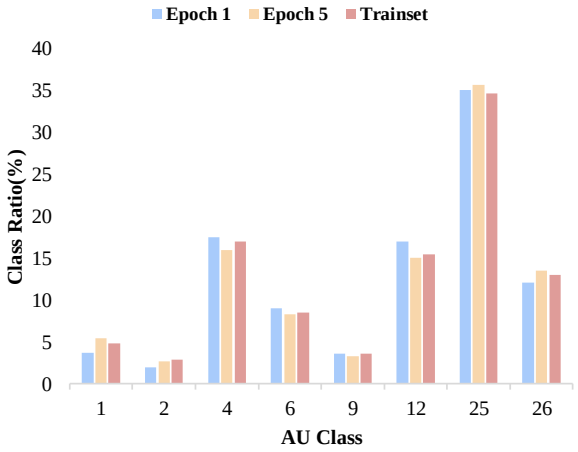

Fig. 4. Visualization of the data distribution estimated in epoch 1, epoch 5, and the entire trainset on DISFA.

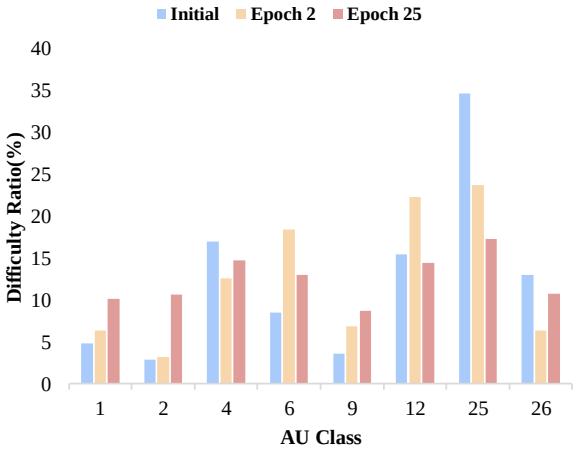

Fig. 5. Visualization of the recognition difficulty (i.e., recall rate) estimated in the initial state, epoch 2, epoch 25 on DISFA. The difficulty ratio is initialized with the data distribution of the entire trainset.

TABLE VII
ABLATION STUDIES OF LOSS FUNCTION ON BP4D

Loss Function	L_{CE}	L_{WA}	L_{DWA}
F1 Score	66.5	66.7	**67.3**

* *L_{CE} refers to the cross entropy loss.*

proposed temporal sampling strategy is more efficient than previous temporal sampling approaches [6]. Our confidence-based AU relationship mining module confirms the state of uncertain AUs while preserving the highly confident AUs, offering an advantage over relation-based methods [13].

D. Ablation studies

We conduct ablation studies on the BP4D dataset to validate each component of our approach. The default setting utilizes the ResNet 50 as the backbone and Cross Entropy loss for the model training. Moreover, We also provide ablation results such as the influence of the sliding window size t and number of subsequences N_{sep}.

Contribution of each component. Table IV evaluates the impact of each component of our framework. It can be observed that introducing the temporal sampling strategy alone increases the F1 score from 62.6% to 65.1% for still image-based AU recognition, demonstrating the effectiveness of our temporal sampling strategy. Additionally, extracting spatial-temporal features using our ST module further boosts the F1 score to 65.7%. Moreover, aligning the direction between the activated AU classes' weights and corresponding spatial features directs the model to focus on regions where AUs are likely to appear, resulting in a further improvement to a 65.8% F1 score. Maintaining consistency between the global and local scores enhances the F1 score to 66.1%. By combining all these components, LAC achieves a promising result of 66.5%, which is comparable to the SOTA method [6], even without the WA loss used by [6]. To address the class imbalance problem, the proposed DWA loss improves LAC to 67.3%, further validating the effectiveness of our loss function.

Analysis of sliding window size t. The size of the sliding window t determines the extent of facial motion that the model can observe each time. As shown in Table V, when t is smaller than 16, the F1 score increases with the growth of t. When t is greater than 16, the F1 score stabilizes within a range of around 65%. We suggest that when t is small, the head motion is negligible, enabling the model to capture AU motions better. As t increases, the motion becomes more prominent, resulting in a higher F1 score. However, when t exceeds 16, the magnitude of facial motion reaches a saturation point. Additionally, the head motion becomes more evident and may confuse the model in distinguishing local facial motions, causing a slight decline in the F1 score. Therefore, we set t to 16 to detect significant facial motions while minimizing the interference of large-scale head motions.

Analysis of number of subsequences N_{sep}. The number of subsequences N_{sep} determines the diversity of the sampled data. In the BP4D dataset, each subject has 8 videos recorded under different visual stimuli, while in the DISFA dataset, only one video is available per subject. To better observe the impact of N_{sep}, we conduct ablation studies on the DISFA dataset. As shown in Table VI, when N_{sep} is less than 12, the F1 score exhibits an upward trend as N_{sep} increases. When N_{sep} is greater than 12, the F1 score decreases slightly. We suggest that when N_{sep} is small, an increase in N_{sep} raises the likelihood of selecting subsequences with different AUs. However, as N_{sep} reaches a certain level, the sampled data may be redundant or without AUs. Moreover, with the random start-frame sampling strategy, an overly large N_{sep} can diminish the sampling diversity, thus resulting in a slight decline in the F1 score. Consequently, we set N_{sep} to 12 to optimize data diversity.

Analysis of DWA loss. As presented in Table VII, we compare the performance of cross-entropy loss (baseline), WA loss, and the proposed DWA loss. The DWA loss demonstrates a superior ability to handle the data imbalance issue. Specifically, it achieves improvements of 0.8% and 0.6% in the F1 score compared to the baseline and WA loss, respectively. We

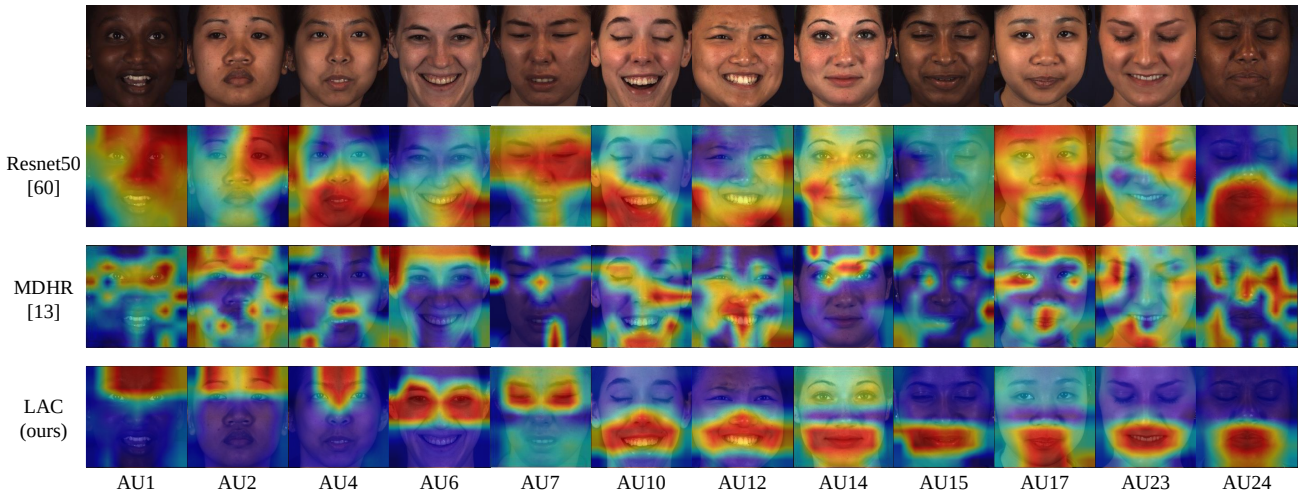

Fig. 6. Visualization of the models' attention, represented by Class Activation Maps (CAM), across different AU classes. The first row is the CAMs estimated by Resnet50 [60], the second row is the CAMs estimated by MDHR [6], and the third row is the CAMs estimated by LAC. It can be observed that LAC localizes AUs more effectively than other methods.

TABLE VIII
ABLATION STUDIES OF MODEL COMPLEXITY

Method	Params (M) ↓	FLOPs (G) ↓
Resnet50 [60]	23.5	66.1
$JÂA$-Net [4]	36.1 (+12.6)	222.3 (+156.2)
MEGraph [13]	31.7 (+8.2)	165.0 (+98.9)
MDHR [6]	93.2 (+69.7)	924.8 (+858.7)
LAC (ours)	**27.2 (+3.7)**	**66.8 (+0.7)**

** FLOPs are estimated when processing a sequence with 16 frames.*

further visualize the two key components that constitute the DWA loss: the data distribution (illustrated in Figure 4) and the difficulty of recognizing AUs (shown in Figure 5). During the training course, the data distribution undergoes subtle changes. The proposed DWA loss detects these alterations and assigns suitable weights to AUs. Moreover, the difficulty of recognizing AUs is the same as the data distribution at the start of the training phase. However, as the training progresses, a misalignment occurs between the recognition difficulty and the data distribution. Specifically, relying solely on data-distribution-based weights is inadequate for facilitating the learning of AUs that are difficult to recognize. Therefore, compared to WA loss, our proposed DWA loss is more effective in modeling the data distribution and promoting the model's learning of each AU.

Analysis of Model complexity. We compare the model complexity of LAC with several previous SOTA methods, including $JÂA$-Net [4], MEGraph [13], and MDHR [6], as presented in Table VIII. Unlike other methods, LAC performs AU recognition without additional complex modules. As demonstrated, LAC requires fewer parameters (i.e., 3.7 M more parameters than ResNet50 and 4.5 M fewer parameters than MEGraph) and has lower FLOPs (i.e., 0.7 G more FLOPs than ResNet50 and 98.2 G fewer FLOPs than MEGraph) in contrast to other methods, highlighting the simplicity and efficiency of our approach.

E. Visualization results

To further validate the feasibility of LAC, we employ Class Activation Maps (CAMs) to visualize the models' attention. Specifically, we select the baseline model (ResNet50 [60]) and the current SOTA method (MDHR [6]) to demonstrate LAC's localization ability. We first visualize CAMs across different AU classes, as shown in Figure 6. It can be observed that ResNet50 can coarsely localize some of the AUs (i.e., AU1, AU2, AU7, AU10, AU15, AU24), but may localize background regions rather than areas where AUs are likely to appear. For example, in the case of AU4 (Brow Lowerer), ResNet50 localizes the lower face, while AU4 should activate in the eyebrow area of the upper face. As for MDHR, the model distributes its attention broadly and fails to localize the precise region. We suggest two possible reasons for this issue: First, MDHR models AU positions coarsely, separating AUs into only three areas. Second, MDHR combines AUs within the same region by recognizing their co-occurrence patterns, which may confuse the model and hinder precise localization. In contrast to ResNet50 and MDHR, LAC explicitly models AU positions, resulting in more accurate localization. Furthermore, we observe that the regions localized by LAC can be distinctly separated into the upper face, middle face, and lower face, as defined in Table I, further validating LAC's effectiveness.

Furthermore, we evaluate LAC's performance on subtle AU detection, as shown in Figure 7. Subtle motion is challenging to detect due to its slight movement and invisible facial patterns. We sample a sequence of frames containing subtle AUs and select AU1 (Inner Brow Raiser) to visualize the models' areas of interest. Additionally, since subtle AUs are difficult to observe from a sequence of images alone, we estimate optical flows to capture the subtle AU motion. To highlight the motion, we calculate the optical flow between non-apex and apex frames. As observed, ResNet50 fails to recognize subtle AU1 and erroneously focuses on the nose. We suggest that this issue can be attributed to still image

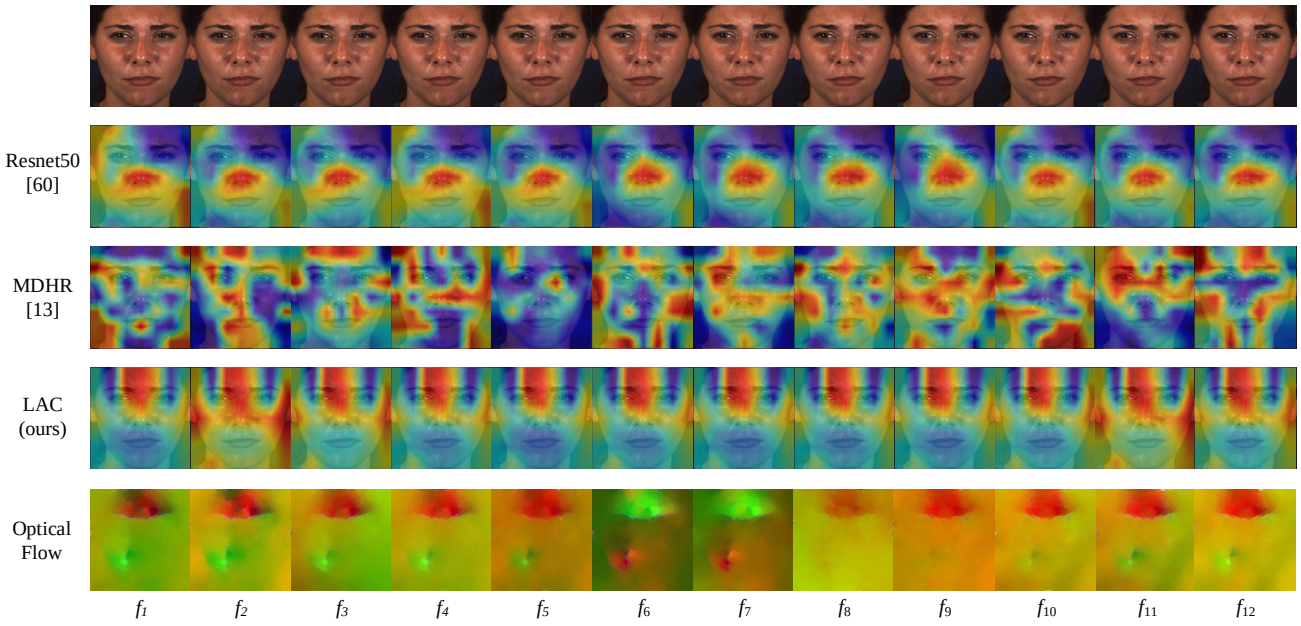

Fig. 7. Visualization of a sequence of CAMs for subtle AU (AU1, Inner Brow Raiser) detection. The first row is the CAMs estimated by Resnet50 [60], the second row is the CAMs estimated by MDHR [6], the third row is the CAMs estimated by LAC, and the fourth row is the optical flow estimated between non-apex and apex frames (i.e., f_6 and f_7) to highlight the subtle motion. It can be observed that LAC localizes subtle motion more precisely than other methods.

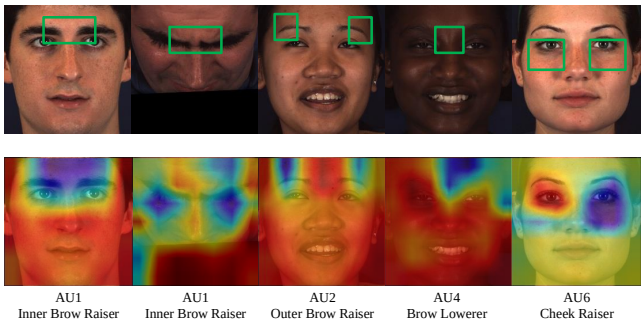

Fig. 8. Visualization of some failure cases of LAC. The first row is the original images, and the green boxes refer to the regions where the corresponding AUs activate. The second row is the CAMs estimated by LAC.

recognition that it cannot detect subtle temporal changes. While MDHR can coarsely localize AU1, its attention map remains diffused. In contrast to them, LAC accurately localizes the inner eyebrow and detects the subtle AU1 motion, further validating the feasibility of our approach.

F. Discussion

As shown in Figure 8, we present some failure cases of LAC where the model localizes the background instead of the corresponding AUs. We suggest that this issue may stem from three potential factors: (i) **Inference sampling method**. During the inference phase, we sample sequences of images with a fixed stride that is equal to the sliding window size. This can lead to situations where the sampled sequence consists of apex frames of subtle AUs, which lack temporal changes and are difficult to detect. (ii) **Face occlusion**. When the facial region is obscured due to head motion or hand movement, LAC fails to localize AUs and accurately identify their states. (iii) **Annotation issues**. We observe that some annotators label AUs on onset frames where no AUs are activated. While LAC can occasionally detect these AUs using temporal information, it remains challenging to recognize them consistently.

In future work, we plan to focus on three key aspects: (i) **Theoretical analysis of temporal sampling strategy**. Although experiments validate the effectiveness of our strategy, we aim to conduct a deeper investigation and provide more rigorous theoretical proof. (ii) **Inference sampling methods**. In this paper, we focus on training sampling strategies; however, we believe there may be more effective methods for sampling data during the inference phase to improve AU recognition performance. (iii) **Strategies to address occlusion**. Facial occlusion is a common issue in real-world scenarios. Therefore, we aim to develop modules or strategies to mitigate this problem, enhancing the model's performance in real-world applications.

V. Conclusion

In this work, we propose a novel framework, Lock and Check (LAC), for AU recognition. Specifically, we decompose AU recognition into two subtasks: AU localization and AU state detection. During the AU localization phase, we design appropriate loss functions and explicitly model AU positions to assist the model in localizing AUs. In the AU detection phase, we introduce a Spatial-Temporal (ST) module to enhance AU features by incorporating AU relationships and temporal information. During the AU relationship mining stage, our confidence-based AU relationship mining module directs the model's focus toward uncertain AUs while maintaining the high-confidence ones. Additionally, the proposed temporal

sampling strategy allows the model to implicitly learn temporal changes, while the DWA loss effectively addresses the data imbalance problem. Empirical evaluations on the BP4D and DISFA datasets demonstrate the superiority of LAC compared to existing methods.

REFERENCES

[1] E. Friesen and P. Ekman, "Facial action coding system: a technique for the measurement of facial movement," *Palo Alto*, vol. 3, no. 2, p. 5, 1978.

[2] K. Zhao, W.-S. Chu, and H. Zhang, "Deep region and multi-label learning for facial action unit detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3391–3399.

[3] Z. Shao, Z. Liu, J. Cai, and L. Ma, "Deep adaptive attention for joint facial action unit detection and face alignment," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 705–720.

[4] ——, "Jaa-net: joint facial action unit detection and face alignment via adaptive attention," *International Journal of Computer Vision*, vol. 129, pp. 321–340, 2021.

[5] J. Yan, J. Wang, Q. Li, C. Wang, and S. Pu, "Weakly supervised regional and temporal learning for facial action unit recognition," *IEEE Transactions on Multimedia*, 2022.

[6] Z. Wang, S. Song, C. Luo, S. Deng, W. Xie, and L. Shen, "Multi-scale dynamic and hierarchical relationship modeling for facial action units recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1270–1280.

[7] C. Corneanu, M. Madadi, and S. Escalera, "Deep structure inference network for facial action unit recognition," in *Proceedings of the european conference on computer vision (ECCV)*, 2018, pp. 298–313.

[8] H. Yang, L. Yin, Y. Zhou, and J. Gu, "Exploiting semantic embedding and visual feature for facial action unit detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 482–10 491.

[9] J. Yu, R. Li, Z. Cai, G. Zhao, G. Xie, J. Zhu, W. Zhu, Q. Ling, L. Wang, C. Wang *et al.*, "Local region perception and relationship learning combined with feature fusion for facial action unit detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5784–5791.

[10] J. Yang, Y. Hristov, J. Shen, Y. Lin, and M. Pantic, "Toward robust facial action units' detection," *Proceedings of the IEEE*, vol. 111, no. 10, pp. 1198–1214, 2023.

[11] Y. Zhang, W. Dong, B.-G. Hu, and Q. Ji, "Classifier learning with prior probabilities for facial action unit recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5108–5116.

[12] S. Wang, B. Pan, S. Wu, and Q. Ji, "Deep facial action unit recognition and intensity estimation from partially labelled data," *IEEE Transactions on Affective Computing*, vol. 12, no. 4, pp. 1018–1030, 2019.

[13] C. Luo, S. Song, W. Xie, L. Shen, and H. Gunes, "Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, 2022, pp. 1239–1246.

[14] Z. Shao, L. Zou, J. Cai, Y. Wu, and L. Ma, "Spatio-temporal relation and attention learning for facial action unit detection," *IEEE Transactions on Image Processing*, vol. PP, 01 2020.

[15] T. Song, L. Chen, W. Zheng, and Q. Ji, "Uncertain graph neural networks for facial action unit detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 7, 2021, pp. 5993–6001.

[16] X. Li, X. Zhang, T. Wang, and L. Yin, "Knowledge-spreader: Learning semi-supervised facial action dynamics by consistifying knowledge granularity," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20 979–20 989.

[17] W.-S. Chu, F. De la Torre, and J. F. Cohn, "Learning spatial and temporal cues for multi-label facial action unit detection," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 2017, pp. 25–32.

[18] W. Li, F. Abtahi, and Z. Zhu, "Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1841–1850.

[19] Z. Wang, S. Song, C. Luo, Y. Zhou, S. Wu, W. Xie, and L. Shen, "Spatial-temporal graph-based au relationship learning for facial action unit detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5899–5907.

[20] L. Yang, I. O. Ertugrul, J. F. Cohn, Z. Hammal, D. Jiang, and H. Sahli, "Facs3d-net: 3d convolution based spatiotemporal representation for action unit detection," in *2019 8th International conference on affective computing and intelligent interaction (ACII)*. IEEE, 2019, pp. 538–544.

[21] N. Churamani, S. Kalkan, and H. Gunes, "Aula-caps: Lifecycle-aware capsule networks for spatio-temporal analysis of facial actions," in *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. IEEE, 2021, pp. 01–08.

[22] G. Li, X. Zhu, Y. Zeng, Q. Wang, and L. Lin, "Semantic relationships guided representation learning for facial action unit recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 8594–8601.

[23] Z. Shao, Z. Liu, J. Cai, Y. Wu, and L. Ma, "Facial action unit detection using attention and relation learning," *IEEE transactions on affective computing*, vol. 13, no. 3, pp. 1274–1289, 2019.

[24] X. Niu, H. Han, S. Shan, and X. Chen, "Multi-label co-regularization for semi-supervised facial action unit recognition," *Advances in neural information processing systems*, vol. 32, 2019.

[25] Y. Chang and S. Wang, "Knowledge-driven self-supervised representation learning for facial action unit recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 417–20 426.

[26] Y. Wei, H. Wang, M. Sun, and J. Liu, "Attention based relation network for facial action units recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[27] Z. Cui, T. Song, Y. Wang, and Q. Ji, "Knowledge augmented deep neural networks for joint facial expression and action unit recognition," *Advances in Neural Information Processing Systems*, vol. 33, pp. 14 338–14 349, 2020.

[28] X. Jia, S. Xu, Y. Zhou, L. Wang, and W. Li, "A novel dual-channel graph convolutional neural network for facial action unit recognition," *Pattern Recognition Letters*, vol. 166, pp. 61–68, 2023.

[29] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=SJU4ayYgl

[30] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, "Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database," *Image and Vision Computing*, vol. 32, no. 10, pp. 692–706, 2014.

[31] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "Disfa: A spontaneous facial action intensity database," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 151–160, 2013.

[32] W. Li, F. Abtahi, Z. Zhu, and L. Yin, "Eac-net: Deep nets with enhancing and cropping for facial action unit detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 11, pp. 2583–2596, 2018.

[33] X. Ge, J. M. Jose, P. Wang, A. Iyer, X. Liu, and H. Han, "Algrnet: Multi-relational adaptive facial action unit modelling for face representation and relevant recognitions," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 5, no. 4, pp. 566–578, 2023.

[34] Y. Chen, D. Chen, Y. Wang, T. Wang, and Y. Liang, "Cafgraph: Context-aware facial multi-graph representation for facial action unit recognition," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 1029–1037.

[35] W. Zhang, L. Li, Y. Ding, W. Chen, Z. Deng, and X. Yu, "Detecting facial action units from global-local fine-grained expressions," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 2, pp. 983–994, 2023.

[36] X. Ge, J. M. Jose, S. Xu, X. Liu, and H. Han, "Mgrr-net: Multi-level graph relational reasoning network for facial action unit detection," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 3, pp. 1–20, 2024.

[37] X. Liu, K. Yuan, X. Niu, J. Shi, Z. Yu, H. Yue, and J. Yang, "Multi-scale promoted self-adjusting correlation learning for facial action unit detection," *IEEE Transactions on Affective Computing*, 2024.

[38] Z. Shao, Y. Zhou, J. Cai, H. Zhu, and R. Yao, "Facial action unit detection via adaptive attention and relation," *IEEE Transactions on Image Processing*, vol. 32, pp. 3354–3366, 2023.

[39] G. M. Jacob and B. Stenger, "Facial action unit detection with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7680–7689.

[40] S. Wang, Y. Chang, and C. Wang, "Dual learning for joint facial landmark detection and action unit recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1404–1416, 2021.

[41] M. Valstar and M. Pantic, "Fully automatic facial action unit detection and temporal analysis," in *2006 conference on computer vision and pattern recognition workshop (CVPRW'06)*. IEEE, 2006, pp. 149–149.
[42] Y. Tong, W. Liao, and Q. Ji, "Facial action unit recognition by exploiting their dynamic and semantic relationships," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 10, pp. 1683–1699, 2007.
[43] Z. Li, Z. Zhang, and L. Yin, "Sat-net: Self-attention and temporal fusion for facial action unit detection," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 5036–5043.
[44] Z. Li, X. Deng, X. Li, and L. Yin, "Integrating semantic and temporal relationships in facial action unit detection," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 5519–5527.
[45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
[46] S. Song, E. Sánchez-Lozano, M. Kumar Tellamekala, L. Shen, A. Johnston, and M. Valstar, "Dynamic facial models for video-based dimensional affect estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.
[47] Z. Cui, C. Kuang, T. Gao, K. Talamadupula, and Q. Ji, "Biomechanics-guided facial action unit detection through force modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8694–8703.
[48] S. He, H. Zhao, J. Juan, Z. Dong, and Z. Tao, "Optical flow fusion synthesis based on adversarial learning from videos for facial action unit detection," in *The International Conference on Image, Vision and Intelligent Systems (ICIVIS 2021)*. Springer, 2022, pp. 561–571.
[49] Z. Shao, Y. Zhou, F. Li, H. Zhu, and B. Liu, "Joint facial action unit recognition and self-supervised optical flow estimation," *Pattern Recognition Letters*, vol. 181, pp. 70–76, 2024.
[50] Y. Li, J. Zeng, S. Shan, and X. Chen, "Self-supervised representation learning from videos for facial action unit detection," in *Proceedings of the IEEE/CVF Conference on Computer vision and pattern recognition*, 2019, pp. 10 924–10 933.
[51] S. Wang, G. Peng, S. Chen, and Q. Ji, "Weakly supervised facial action unit recognition with domain knowledge," *IEEE transactions on cybernetics*, vol. 48, no. 11, pp. 3265–3276, 2018.
[52] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proceedings of the national academy of sciences*, vol. 111, no. 15, pp. E1454–E1462, 2014.
[53] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=rJXMpikCZ
[54] Y. Chen, C. Chen, X. Luo, J. Huang, X.-S. Hua, T. Wang, and Y. Liang, "Pursuing knowledge consistency: Supervised hierarchical contrastive learning for facial action unit recognition," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 111–119.
[55] Z. Zhou, H. Li, H. Liu, N. Wang, G. Yu, and R. Ji, "Star loss: Reducing semantic ambiguity in facial landmark detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 15 475–15 484.
[56] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.
[57] Y. Yu, S. Khadivi, and J. Xu, "Can data diversity enhance learning generalization?" in *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 4933–4945. [Online]. Available: https://aclanthology.org/2022.coling-1.437/
[58] X. Li, Z. Zhang, X. Zhang, T. Wang, Z. Li, H. Yang, U. Ciftci, Q. Ji, J. Cohn, and L. Yin, "Disagreement matters: Exploring internal diversification for redundant attention in generic facial action analysis," *IEEE Transactions on Affective Computing*, vol. 15, no. 2, pp. 620–631, 2023.
[59] Y. Chang, C. Zhang, Y. Wu, and S. Wang, "Facial action unit recognition enhanced by text descriptions of facs," *IEEE Transactions on Affective Computing*, 2024.
[60] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[61] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

Zihao Huang received the B.S. degree in computer science and technology from South China University of Technology, Guangzhou, China in 2022. He is currently the postgraduate working towards Ph.D. degree of computer science and technology in South China University of Technology, Guangzhou, China. His research interests focus on computer vision, including image processing, affective computing, facial action unit recognition, etc.

Jian Gao received B.S. degree in Internet of Things Engineering from Jilin University (2021). He is currently an M.S. candidate in Computer Science and Technology at South China University of Technology. His research specializes in computer vision-based affective computing, particularly facial action unit analysis.

Wentian Cai is currently working toward the Ph.D. degree of computer science and technology in the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. His research interests include computer vision, deep learning, weakly supervised learning, and medical image analysis.

Yandan Chen received the B.S. degree in computer science and technology from South China University of Technology, Guangzhou, China. She is currently working towards Ph.D. degree of computer science and technology in South China University of Technology, Guangzhou, China. Her research interests include image processing and deep learning.

Xiping Hu received the Ph.D. degree from the University of British Columbia, Vancouver, BC, Canada. He is currently a professor with Beijing Institute of Technology, and with Shenzhen MSU-BIT University, China. He has more than 150 papers published and presented in prestigious conferences and journals, such as IEEE TPAMI/TMC/TPDS/TIP/JSAC, IEEE COMST, ACM MobiCom/MM/SIGIR/WWW, AAAI, and IJCAI. He has been serving as associate editor of IEEE TCSS, and the lead guest editors of IEEE IoT Journal and IEEE TASE etc. He has been granted several key research projects with more than 50,000,000 RMB as principal investigator. He was the Co-Founder and CTO of Bravolol Ltd., Hong Kong, a leading language learning mobile application company with over 100 million users, and listed as the top 2 language education platform globally. His research areas consist of mobile cyber-physical systems, crowdsensing and affective computing.

Ping Gao received the master's and doctor's degrees from XiangYa School of Medicine, Central South University, China, in 1998 and 2002, respectively. She is currently a chief physician at Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences. She has published more than 20 papers. Her research interests include respiratory diseases and medical intelligence. She participated in the treatment of critically ill COVID-19 patients.

Ying Gao received the Bachelor's degree, Master's degree in computer science from Central South University of China and the Ph.D. degree in computer science from South China University of Technology, China, in 1997, 2000 and 2006, respectively. She is currently a professor with the School of Computer Science and Engineering, South China University of Technology, China. Her current research interests include computer vision, deep learning, and network security.