Applied Intelligence
https://doi.org/10.1007/s10489-023-05154-7

# Learning facial expression-aware global-to-local representation for robust action unit detection

**Rudong An[1] · Aobo Jin[2] · Wei Chen[3] · Wei Zhang[1] · Hao Zeng[1] · Zhigang Deng[4] · Yu Ding[1]**

Accepted: 2 November 2023


**Abstract**
The task of detecting facial action units (AU) often utilizes discrete expression categories, such as Angry, Disgust, and Happy, as auxiliary information to enhance performance. However, these categories are unable to capture the subtle transformations of AUs. Additionally, existing works suffer from overfitting due to the limited availability of AU datasets. This paper proposes a novel fine-grained global expression representation encoder to capture continuous and subtle global facial expressions and improve AU detection. The facial expression representation effectively reduces overfitting by isolating facial expressions from other factors such as identity, background, head pose, and illumination. To further address overfitting, a local AU features module transforms the global expression representation into local facial features for each AU. Finally, the local AU features are fed into an AU classifier to determine the occurrence of each AU. Our proposed method outperforms previous works and achieves state-of-the-art performances on both in-the-lab and in-the-wild datasets. This is in contrast to most existing works that only focus on in-the-lab datasets. Our method specifically addresses the issue of overfitting from limited data, which contributes to its superior performance.

**Keywords** Facial action coding · Facial action unit detection · Facial expression recognition · Expression-aware representation · Deep learning

## 1 Introduction

Facial Affect Analysis is a vibrant and rapidly-evolving research field within the computer vision and affective computing communities. It encompasses two primary descriptors: facial expression and Facial Action Units (AUs). As a universal non-verbal method of communication, facial expressions can be effectively analyzed by combining various AUs, as determined by the Facial Action Coding System (FACS) [6]. These AUs include 32 atomic facial action descriptors, which are based on distinct anatomical facial muscle groups. By utilizing FACS and AUs, researchers can gain valuable insights into human emotions and behavior, fostering the development of more accurate and robust facial affect analysis systems. Each AU characterizes the movement of a specific facial region. For instance, the Inner Brow Raiser (AU1) descriptor focuses on the medial portion of the frontalis muscle. AUs and facial expressions are inextricably linked, as nearly any conceivable expression can be accurately described as a specific combination of facial AUs. For instance, as illustrated in Fig. 1, a happy expression can be achieved by the occurrence of both AU6 and AU12; a doubtful expression is related to AU4, and a surprised expression is linked to AU1 and AU26. In this regard, AU features are frequently regarded as a facial expression representation. Undeniably, AUs play a pivotal role in conveying human emotions and enabling automatic expression analysis. The detection of AUs has garnered significant attention in recent years, owing to its vast array of applications, such as emotion recognition [26], micro-expression detection [43], talking face generation [19], and mental health diagnosis [28].

The detection of AUs poses a significant challenge due to the prevalence of overfitting resulting from the scarcity of annotated data. Even for well-trained experts, manually labeling a hundred images with AUs requires hours of tedious effort [53]. Consequently, acquiring well-annotated AUs data on a large scale is both time-consuming and laborious [37]. On one hand, this limits the effectiveness of deep networks due to data overfitting and severe data imbalance. On the

✉ Yu Ding
dingyu01@corp.netease.com

Extended author information available on the last page of the article

Published online: 06 January 2024

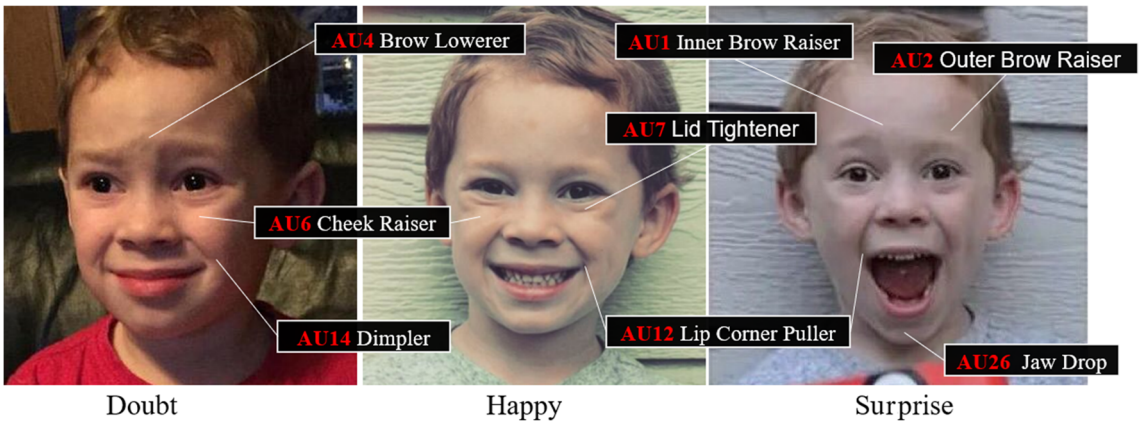



**Fig. 1** Examples of three expressions and their corresponding AUs. From left to right: doubt, happy, surprise. Expression is a global description of the face muscle movements, while AUs refer to the individual local description of muscle motion

other hand, AU datasets often feature a limited number of identities, ranging from dozens [22, 49] to hundreds [50], leading to identity overfitting [24, 39]. This paper aims to tackle the aforementioned overfitting challenges.

**Global learning** Compared to AU datasets, there are several readily accessible and easy-to-annotate expression datasets, such as Facial Expression Coding System (FEC) [36] and AffectNet [23]. These datasets are useful in improving AU detection performance due to their close correlations [39]. Expression recognition has also been explored to enhance AU recognition tasks [5, 18]. However, considering expressions as just a few rough discrete classes is sub-optimal for learning subtle expression transformations [5, 18].

In contrast, this paper aims to capture the subtle distinctions between expressions through similarity learning, by utilizing continuous and compact expression features for AU representations. Additionally, similarity annotation can be easily performed by non-professional participants, without requiring certified AU annotators. Large-scale annotation efforts aid in disentangling identity as much as possible [48].

**Local learning** Facial AUs are defined anatomically based on the movements of their corresponding facial muscles, and therefore, they are inherently related to local facial regions, referred to as AU's local characteristics. The exploration of these local characteristics of AUs is also considered as a potential solution to alleviate overfitting.

Effectively extracting local features from an input face image is a crucial yet unresolved challenge in this regard, as local facial deformations are not only subtle and transient [53], but they also vary among individuals [24, 39]. The conventional approach involves dividing the input image into several local parts in a predefined manner and feeding them into networks to capture local information, commonly known as patch learning [25, 54]. Such methods pre-define AU local regions with prior knowledge and crop a face image into several regular yet coarse-grained patches. However, this approach is limited by the quality of cropped patches, which can negatively impact the feature extractor's performance.

Another category of approaches involves leveraging structured geometric information, such as landmarks, to generate more detailed and fine-grained local partitions, such as Regions of Interest (ROIs) [51] or attention maps [13, 17, 31, 32, 44]. Recent research studies [13, 17, 32] have emphasized learning attention maps in a supervised manner by utilizing facial landmarks. During training, the attention maps are often initialized with pre-defined AU centers based on landmarks and then treated as the ground truth.

Nevertheless, facial AUs are usually not restricted to landmark points but occur in specific locations [21]. Therefore, pre-defined AU regions or centers in a uniform manner, whether based on landmarks or prior domain knowledge, are prone to errors. For instance, several prior AU patch generation techniques are fixed for different head positions, whereas landmark detection often deviates under significant head rotations, leading to imprecise AU patches or attention [25]. Overall, these approaches have two major drawbacks. Firstly, AU centers, which are based on domain knowledge and landmarks, are restricted to landmark positions and/or predetermined rules [21, 31]. Secondly, errors in landmark detection frequently impact the final performance [21, 31]. Therefore, relying on manually pre-defined rules with coarse guidance to capture local features could limit the potential of local feature extractors. Consequently, the precise position of each AU in the facial feature maps ought to depend on the specific AU detection task and the training dataset. In other words, the learning process of the attention maps should be driven by the data.

**Overview** Inspired by the aforementioned observations, we present a pioneering framework for AU detection called the

*Global-to-Local Expression-aware Network (GTLE-Net)*, which aims to overcome the aforementioned challenges. Firstly, to address the challenges of data and identity overfitting, we pre-train a *global facial expression representation encoder (GEE)* on a large-scale facial expression dataset, FEC [36], to extract identity-invariant global expression features. Rather than classifying expressions into discrete categories, we project them into a *continuous* representation space, which proves to be more effective for AU detection in our experiments given that AUs are inherently subtle and continuous. Secondly, to capture local features, we design a *local AU feature module (LAM)* with two extractors that produce AU masks and feature maps for each AU, respectively. Notably, the AU mask extractor is learned without any intermediate supervision, meaning that the attention solely depends on the input data. Finally, we obtain *AU masked features* by multiplying the AU feature maps and the corresponding AU masks, which effectively filter out irrelevant fields and retain informative ones. Our framework has the distinct advantage of utilizing large-scale, well-annotated expression data to its fullest potential. Furthermore, it can learn attention maps adaptively without requiring any extra supervision, which effectively enhances AU detection.

To sum up, the main contributions of this paper can be summarized below.

1. This paper presents a novel facial action unit detection framework that intentionally targets the issue of overfitting, which is commonly encountered with limited training data. Through extensive experimentation, the validity and accuracy of our proposed method have been fully established across four commonly used benchmark datasets. In fact, our method has proven to surpass the current state-of-the-art solutions by a significant margin.
2. Our proposed solution relies on jointly learning intermediate latent spaces of both a global facial expression and local AU-specific facial regions. The global facial expression is based on a pretraining model of facial expression similarity and it is used to produce the local AU-specific latent spaces.

The remainder of this paper is organized as follows. Related works are presented in Section 2. The details of our method are described in Section 3. Comprehensive experiment results and a series of visual analyses are reported in Section 4. Discussion and concluding remarks are provided in Section 5. This paper has a preprint version on arxiv[1] that will be updated after the acceptance of this paper.

[1] https://arxiv.org/pdf/2210.15160v2.pdf

## 2 Related work

Over the past few years, researchers have explored numerous approaches for detecting AUs, including those incorporating multi-task and attention mechanisms, as well as those integrating auxiliary information such as landmarks, text descriptions, and facial expressions. In the following section, we will provide an overview of previous related works.

### 2.1 AU detection with auxiliary information

Due to the high cost of annotating AUs, existing AU detection datasets are often limited in scale and subject variation. Consequently, previous methods for AU detection have relied on various forms of auxiliary information to improve generalization performance. Among these features, facial landmarks are the most frequently used pre-trained features for AU detection. For instance, EAC-Net [17] constructs local regions of interest and spatial attention maps from facial landmarks. LP-Net [24] trains a person-specific shape regularization network from detected facial landmarks, which reduces the influence of person-specific shape information and yields more AU-discriminative features. JAA-Net [30] is the first work to jointly perform AU detection and facial landmark detection from data annotated with both labels. These works demonstrate the efficacy of facial landmarks as auxiliary information in AU detection. However, since these methods rely on landmark detection results as prior knowledge and the locations of AUs are confined to a pre-defined set of positions [31], their detection results may be affected by the performance of landmark detection.

Several other types of auxiliary information have been investigated for improving AU detection. Zhao et al. [52] pre-trained a weakly supervised embedding using a large dataset of web images. Cui et al. [5] utilized prior probabilities of AU occurrences as generic knowledge to jointly optimize emotions and AUs. ME-GeaphAU [20] proposed a novel approach that models the relationship between each pair of AUs explicitly via node and edge features in a graph model, achieving superior performance. WSRTL [40] introduced RoI inpainting and optical flow estimation as auxiliary tasks, outperforming existing methods. Recently, SEV-Net [44] leveraged textual descriptions of AU occurrences by utilizing pre-trained word embedding to obtain auxiliary textual features. These studies demonstrated the effectiveness of incorporating various types of auxiliary information in AU detection, such as unlabeled data, prior probabilities of AU occurrences, emotions, inpainting, optical flow, and textual descriptions. However, the effectiveness of compact expression embedding as auxiliary information has not been explored. In this paper, we introduce pre-trained expression embedding as auxiliary information to enhance the generalization of our model.

### 2.2 AU feature learning

Due to the variations in local definitions of AUs, it is crucial to extract local AU features. To achieve this, some researchers propose patch learning to obtain local information. For example, Zhong et al. [54] preprocess input images into uniform patches before encoding them to analyze facial expressions. Onal et al. [7, 25] consider head pose and crop AU-specific local facial patches containing information for specific AU recognition after registering the 3D head pose to reduce the impact of head movements. Additionally, attention mechanisms are commonly used to highlight features at facial AU-based positions. Facial landmarks with sparse facial geometric features are advantageous as a supervised attention prior. EAC-Net [17] creates fixed attention maps related to the correlations between AUs and landmarks. JÂA-Net [32] performs joint AU detection and facial landmark detection, and the predicted landmarks are used to compute the attention map for each AU. Jacob et al. [13] propose a multi-task approach that combines AU detection and landmark-based attention map prediction. ARL et al. [31] proposed a channel-wise and spatial attention learning for each AU and a pixel-level relation is learned by CRF to refine the spatial attention. Except for the facial landmarks, the difference saliency map [1] is also validated to be effective in feature learning for AU detection. SEV-Net [44] utilizes the textual descriptions of local details to generate a regional attention map. In this way, it highlights the local parts of global features. However, it requires extra annotations for descriptions. Our work proposes a pixel-wise attention map that is data-driven without explicit supervision. The experiments show that our method is able to generate attention maps adaptively and effectively in promoting AU detection.

### 2.3 Expression representations

Facial expressions are crucial for accurate AU detection. Some previous studies have utilized large amounts of available expression data to improve the performance of AU detection [5, 52]. Various approaches have been proposed to transform face images into a lower-dimensional manifold for subject-independent expression representations. However, earlier works [23] trained embeddings for discrete emotion classification tasks that neglect the facial expression variations within each class.

The 3D Morphable Model (3DMM) [27] has been developed to fit identities and expression parameters from a single face image. In 3DMM, expressions are represented as the coefficients of predefined blendshapes. These estimated expression coefficients are then utilized for talking head synthesis [16], expression transfer [46], and face manipulation [8]. However, the estimated expression coefficients exhibit a limitation in their ability to represent fine-grained expressions.

To address this limitation, Vemulapalli and Agarwala [36] proposed a solution in the form of a compact facial expression embedding that captures subtle and complex expressions. This embedding defines facial expression similarity through triplet annotations. Similarly, Zhang et al. [48] proposed a Deviation Learning Network (DLN) to generate more compact and smooth representations of facial expressions by removing identity information from continuous expression embeddings. While these works demonstrate the promising potential of compact and continuous expression embeddings, their impact on AU detection is unknown. This motivated us to explore the use of such embeddings to enhance AU detection, which can be seen as an extension of our previous work [48].

## 3 Proposed method

This section provides a brief overview of the problem definition and introduces the proposed GTLE-Net framework. GTLE-Net is a novel architecture designed to extract both global expression and local AU features to address the issue of overfitting. The global expression component is designed to leverage fine-grained representation and disentangle the identity component from the expression, which can alleviate the overfitting of limited AU annotation and identities. Additionally, the local AU features are also incorporated to help alleviate the issue of limited AU annotation. By combining both global and local features, GTLE-Net aims to provide a more robust and accurate approach to AU detection.

### 3.1 Problem definition

The task of AU detection involves predicting the occurrence probabilities of AUs, denoted by $[o^1, o^2, ..., o^N]$, given a face image with a resolution of $256 \times 256$. As discussed in Section 1, global and local expressions are two distinct levels of description that are highly interrelated and can be leveraged to enhance each other. While facial expression data are readily available and annotated, AUs are more subtle and challenging to annotate. However, extracting facial AU information from facial expression representations becomes feasible when expression representations are fine-grained. Therefore, we propose a novel architecture, GTLE-Net, which includes a *global expression encoder (GEE)* pretrained on a large-scale facial expression dataset to obtain a robust expression representation. Additionally, we propose a *local AU features module (LAM)* consisting of two extractors, namely the AU mask extractor ($E_m$) and the AU feature map extractor ($E_a$), to capture local AU features. The frame-

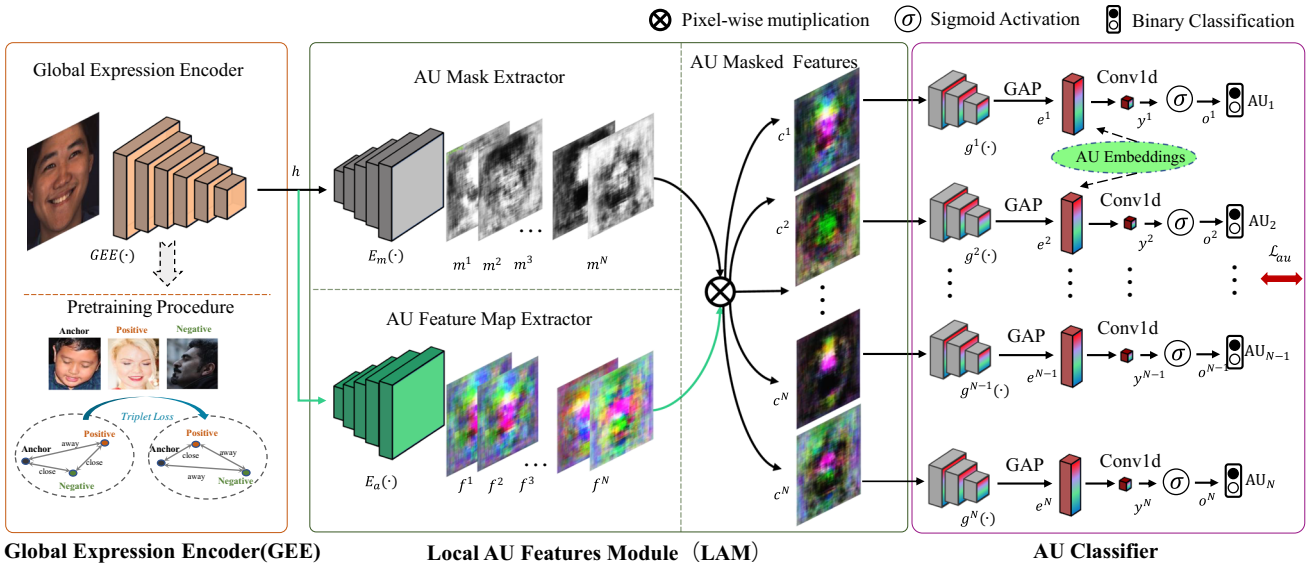



**Fig. 2** The schematic pipeline of the proposed global-to-local expression-aware representation AU detection framework. It is composed of three main modules: the Global Expression representation Encoder (*GEE*), the local AU features module (*LAM*), and the AU classifier. The *GEE*, pre-trained by expression similarity tasks with the triplet loss, extracts a global expression feature map from the input image, which is capable of sensing subtle expression changes. The global expression feature is then fed into the AU Mask Extractor ($E_m$) and the AU Feature Map Extractor ($E_a$). The two extractors produce AU-specific masks and feature maps from the global expression feature ($h$), respectively. AU masked features are obtained through pixel-wise product on the outputs of two extractors, which are then fed into the following AU Classifier to obtain the AU embeddings and detection results

work is illustrated in Fig. 2. In the following, we provide a detailed discussion of each main module.

## 3.2 Global expression encoder

As previously mentioned, the availability of datasets is limited, which may result in overfitting to unexpected factors such as identity, head pose, background, and so on [48]. To improve the generalization of our model, we incorporate a compact and continuous expression embedding as prior auxiliary information. This is achieved by pre-training our model, known as *GEE*, on the expression similarity task using the FEC dataset [36]. The FEC dataset consists of numerous facial image triplets, each of which includes annotations indicating which image has the most different expression. By pre-training *GEE* using the triplet loss, we constrained the embedding distance of more similar expressions to be smaller. In each triplet, we refer to the image with the most different expression as the Negative ($N$), and the other two images as the Anchor ($A$) and Positive ($P$), respectively. For a triplet (A, P, N), $GEE$ outputs three expression embeddings $E_A$, $E_P$ and $E_N$. The pre-training process involves using the triplet loss in the following manner:

$$\begin{aligned}\mathcal{L}_{tri} = &\max(0, \|E_A - E_P\|_2^2 - \|E_A - E_N\|_2^2 + m) \\ &+ \max(0, \|E_A - E_P\|_2^2 - \|E_P - E_N\|_2^2 + m).\end{aligned} \tag{1}$$

The triplet loss employed in the pre-training process enables *GEE* to learn an expression embedding space that maps similar expressions closer and different ones farther away. With the large number of triplets present in the training data, the pre-trained feature encoder is capable of generating a compact and identity-invariant expression representation. This is highly advantageous for downstream expression-related tasks such as AU detection.

We trained *GEE* with four commonly used candidate backbones and determined that InceptionResnetV1 [34] produced the best performance (refer to Table 1). Consequently,

**Table 1** Ablation study on GEE frameworks by Accuracy (%). The reported results are the triplet accuracy (ACC %) on the FEC validation set, with different frameworks trained on the FEC training dataset

| Framework | VGG16 | ResNet50 | RepVGGA2 | InceptionResNetV1 |
|---|---|---|---|---|
| ACC(%) | 81.12 | 81.13 | 81.49 | **86.10** |

The best results are shown in bold

we selected it as the backbone of *GEE* and replaced the last classifier layer with a linear layer, which was subsequently followed by normalization. After dimensionality reduction, it outputs a 16-dimensional vector as the expression embedding. Our pre-training process involved training *GEE* on the expression similarity task. We then utilized the shallow layers (up to the layer with feature maps in a resolution of $16 \times 16$) to extract the global expression representation $h$ from the input face image $im$. This process is illustrated by the following equation:

$$h = GEE(im) \tag{2}$$

Then, our whole framework including $GEE$ is fine-tuned for AU detection.

We have two main intuitions. Firstly, we believe that initializing *GEE* with expression embedding learning will enable the AU detection framework to capture more effective expression features. Secondly, it provides a strong initialization at the outset of training, thereby minimizing the problem of over-fitting, which is commonly associated with AU datasets due to their limited data and identities.

### 3.3 Local AU features module

The expression features extracted by the pre-trained $GEE$ capture facial movements globally, while disentangling them from redundant factors such as identity, head pose, background, hair, and so on (please refer to Fig 6 and 7). This helps to prevent overfitting to these factors, which can be particularly problematic given the limited data available. Moreover, since AUs are characterized by subtle, localized motions in particular facial regions, it is crucial to learn local representations to effectively detect each AU.

To achieve this goal, we design the Local AU Features Module ($LAM$), which consists of two parallel extractors: an AU mask extractor ($E_m$) and an AU feature map extractor ($E_a$), to transform the global expression features into local AU-related representations.

Our approach involves extracting AU features ($f^i$ $(i = 1, 2, ..., N)$) using the AU feature map extractor, $E_a$, from the global expression features. These AU features are distinct to each specific AU. However, we identified that $E_a$ may generate AU features containing information from irrelevant AUs. To resolve this, we introduce an AU mask extractor ($E_m$) to locate the region of interest for each AU and eliminate unrelated content. In this way, we ensure that the extracted AU features are precise and accurately represent each specific AU.

**AU feature map extractor** The AU feature map extractor, $E_a$, is designed to generate AU-related features. To accomplish this, $E_a$ utilizes the global expression feature, $h$, and operates as follows:

$$[f^1, f^2, ..., f^N] = E_a(h), \tag{3}$$

The $k$-th AU feature map denoted as $f^k \in \mathbb{R}^{3\times H\times W}$, corresponds to the $k$-th AU, where $N$ represents the total number of all AUs. Due to the abundance of AUs on the face, designing an individual extraction branch for each AU will result in a large number of parameters. Thus, we propose a shared backbone for the AU feature map extractor with multiple prediction heads. As illustrated in Fig. 3(a), the architecture of $E_a$ is derived from the generator in StarGAN v2 [4], with a switch from AdaIN [12] to instance normalization (IN) [35]. The shared backbone comprises stacked residual blocks [10] and upsampling layers, which magnify local facial details to benefit the subsequent AU detection task. Each prediction head consists of a convolutional layer and an activation layer and is responsible for predicting the facial content related to a specific AU.

**AU mask extractor** The task of the AU mask extractor $E_m$ is to create AU mask maps based on the global expression features $h$, with each mask indicating the specific region of interest for a particular AU.

$$[m^1, m^2, ..., m^N] = E_m(h), \tag{4}$$

where $m^k \in \mathbb{R}^{1\times H\times W}$ is the $k$-th facial mask for the $k$-th AU.

As illustrated in Fig. 3(b), the architecture of $E_m$ is similar to $E_a$. In order to constrain the predicted facial masks to a standardized range of 0 to 1 (where 0 signifies "no attention" and 1 represents "strongest attention"), the final layer of $E_m$ is augmented with an activation function. While Sigmoid and Softmax both serve this purpose, Softmax often produces sparse attention maps that disregard regions relevant to the target AU. Consequently, the downstream recognition task may not receive sufficient information to learn effectively. As a result, we opt to use Sigmoid as the activating function for the outputs of $E_m$.

**AU masked features** After obtaining the AU-specific feature maps and masks, we generate the final AU masked features by performing a multiplication operation between them:

$$[c^1, ..., c^N] = [m^1 \times f^1, ..., m^N \times f^N], \tag{5}$$

The computed masked AU feature for the $i$-th AU is denoted as $c^i$, where $i$ represents an integer in the range $(1, 2, ..., N)$. These masked AU features are subsequently used in the AU detection task. In this way, the $LAM$ algorithm facilitates the generation of these masked features, which are then passed on to an AU Classifier (described below) for the recognition of AUs.

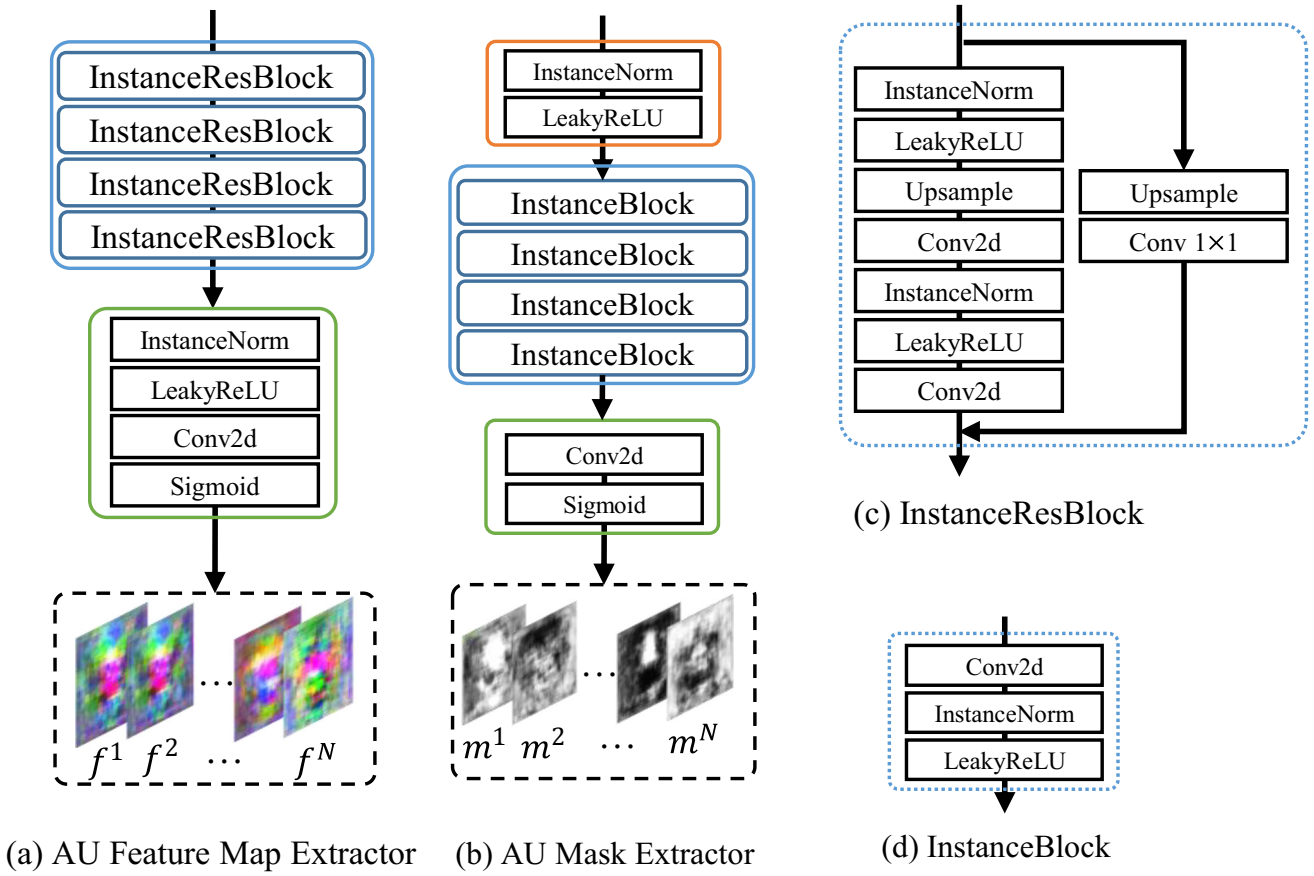



**Fig. 3** The architectures of (a) AU Feature Map Extractor and (b) AU Mask Extractor. Their main blocks consist of several (c) Instance Norm Residual Blocks and (d) Instance Norm Blocks, respectively. Their output modules consist of convolution layers with Sigmoid activation that output AU feature maps and masks

### 3.4 AU classifier

The $LAM$ algorithm generates a set of RGB feature maps with high resolution ($256 \times 256$) as its output. Since the number of these feature maps corresponds to the number of AUs in the given dataset, the subsequent AU classifier module, denoted as $g^i(\cdot)$, is designed as a multi-layer CNN architecture comprising 6 convolution blocks and a global average pooling layer. The first five convolution blocks are followed by batch-norm and ELU activation functions, and their output channels are 16, 32, 64, 128, and 256, respectively. To ensure a broad receptive field, the kernel size, stride, and padding of each of these convolution blocks are set to 7, 2, and 3, respectively. The global average pooling layer is positioned after the fifth convolution block to aggregate the feature maps into AU Embeddings, which extract more information relevant to the specific AU from the input AU masked feature map. The AU embeddings are then input into the last convolution layers, where the output channel, kernel size, stride, and padding are set to 1. This process can be represented mathematically as the following function:

$$[y^1, y^2, ..., y^N] = [g^1(c^1), g^2(c^2), ..., g^N(c^N)]. \quad (6)$$

The output of the last convolution layer of the $i$-th AU is represented by $y^i$. The predicted probability $o^i$ of the AU's probability is obtained by applying a Sigmoid activation on $y^i$, which can be expressed as follows:

$$[o^1, o^2, ..., o^N] = [\sigma(y^1), \sigma(y^2), ..., \sigma(y^N)]. \quad (7)$$

### 3.5 Loss functions

Our proposed framework consists of two steps, as shown in Fig. 2. The first step involves the pretraining procedure, which is focused on learning expression similarity using the triplet loss ($L_{tri}$), as defined in Equation (1). In the second step, the problem of AU detection is treated as a multi-label binary classification problem, and the AU detection loss is

defined accordingly:

$$\mathcal{L}_{au} = -\sum_{i=1}^{N}[p_i \log \hat{p}_i + (1 - p_i) \log(1 - \hat{p}_i)], \tag{8}$$

where the number of AU classes is represented by $N$, and $p_i$, $\hat{p}_i$ denote the ground truth and predicted probability, respectively, for the $i$-th AU.

## 4 Experimental results

In this section, we present a comprehensive set of experiments to demonstrate the effectiveness of our proposed method on in-the-lab and in-the-wild datasets. The subsequent sections are structured as follows.

We first provide a detailed description of the four datasets used in our study, followed by an explanation of the implementation details of the training and validation procedures.

Furthermore, we conduct a comparative analysis to demonstrate the superior performance of our proposed approach. To facilitate a comprehensive evaluation, we report F1-score, accuracy, and AUC results on each dataset. To ensure fair comparisons, we adopt the evaluation settings used in prior works [17, 32, 42]. Specifically, for the three in-the-lab datasets (BP4D, BP4D+, and DISFA), we employ a three-fold cross-validation approach and follow the specific division of training and validation data as outlined in [17, 32]. The reported results are an average of the three-fold experiments conducted on each dataset. For the in-the-wild dataset (RAF-AU), we adopt the evaluation and dataset split approach used in [42], where the training section is four times larger than the test section.

Thirdly, to further obtain insight into our method, we design ablation experiments to study the impact of the built sub-modules, including the proposed global expression representation encoder and local AU features module.

Finally, to evaluate the robustness and generalization capability of our model, we conduct a cross-dataset evaluation study and a partial dataset experiment. Additionally, we analyze the training and validation losses and visualize several samples to validate the effectiveness of our proposed approach.

### 4.1 Datasets

As previously mentioned, our evaluation of the proposed approach was conducted on four widely-used AU detection datasets: BP4D [49], DISFA [22], BP4D+ [50] and RAF-AU [42]. Each of these datasets was annotated with AU labels by certified experts for each frame. The order of these datasets from largest to smallest is BP4D+, BP4D, DISFA, and RAF-AU. While BP4D+, BP4D, and DISFA contain in-the-lab data, RAF-AU records in-the-wild data. BP4D and BP4D+ were recorded in similar shooting environments and light conditions, and both share the same 12 AU classes. In comparison, DISFA is a smaller dataset containing only 8 AU classes and was captured in quite different lighting conditions. Notably, RAF-AU is the smallest dataset with 13 AU classes, but it consists of complex facial expression images in the wild.

**BP4D** contains 328 video clips featuring 41 participants, consisting of 23 females and 18 males. Each frame of the videos was annotated with the presence or absence of 12 AU classes (1, 2, 4, 6, 7, 10, 12, 14, 15, 17, 23, and 24), totaling around 140,000 annotated frames by certified FACS experts. For consistency with previous studies [17, 32], we employed a three-fold, subject-exclusive, cross-validation technique for all experiments.

**DISFA** comprises 27 video clips featuring 27 participants, including 12 females and 15 males. Each clip captures the facial activity of a participant while watching a 4-minute video, consisting of 4,845 evenly spaced frames. A total of approximately 130,000 frames were annotated with intensity levels ranging from 0 to 5, with labels of {0, 1} indicating absence and all others indicating presence, as established in prior research [13, 17, 32, 44]. We focused our evaluation on 8 AU classes (1, 2, 4, 6, 9, 12, 25, and 26), using the same subject-exclusive, three-fold cross-validation methodology as previous studies [17, 32].

**BP4D+** includes 1400 video clips featuring 140 participants (82 females and 58 males), presenting greater identity variations and video numbers than BP4D. It is worth mentioning that both datasets share the same 12 AUs, and around 198,000 frames were annotated in BP4D+. Our study employed a three-fold, subject-exclusive, cross-validation methodology to compare our approach with state-of-the-art methods. Additionally, to evaluate the generalization performance, we trained our model on BP4D and tested it on BP4D+, as previously done in other studies [31, 32].

**RAF-AU** is a facial expression dataset with AU annotations that captures real-world scenarios. It comprises 4601 highly diverse facial expression images, along with annotations for 26 different AUs. The distinguishing feature of RAF-AU is that it includes images with complex expressions, occlusions, variations in illumination, resolution, and head poses that present a challenge to deep learning models. To facilitate validation, we follow the baseline in [42] and evaluate and analyze 13 AUs (AU 1, 2, 4, 5, 6, 9, 10, 12, 16, 17, 25, 26, and 27).

### 4.2 Implementation details

The face images are first aligned and cropped based on landmark detection results. Before feeding to the networks, we resized the input images to 256 × 256. After passing through $GEE$, the resolution of the resulting feature maps was reduced to 16 × 16.

To enhance the feature extraction capability of low-level convolution layers, $GEE$ was initialized with the corresponding network parameters pretrained on the dataset of ImageNet or FEC, and the rest of our networks were set to Kaiming initialization [9]. Except for the AU Mask Extractor ($E_m$), the whole model parameters were updated by employing an Adam optimizer with hyper-parameters $\beta = 0.9$ and weight decay $10^{-6}$.

For each dataset fold, the networks underwent 20 epochs of training with an initial learning rate of $10^{-5}$, using a mini-batch size of 10. A warm-up stage of $10^3$ steps was employed, followed by an exponential decay of the learning rate with an exponent of $-0.5$. The training and inference pipelines were implemented in PyTorch and utilized 4 GeForce GTX 2080ti GPUs.

### 4.3 Comparisons with the state-of-the-art

We compared our method with many state-of-the-art AU detection methods, including HSTR-Net [33], MMA-Net [29], GeoCNN [2], DC-Net [11], EAC-Net [17], ARL [31], SRERL [15], ML-GCN [3], MS-CAM [47], FACS3D-Net [45], JÂA-Net [32], SEV-Net [44], D-PAttNet$^{tt}$[25], ME-GraphAU [20], Dual-Learning [38], WSRTL [41], and AU-CNN[42]. Except for AU-CNN validated with the in-the-wild dataset of RAF-AU, all the above other ones are validated only with in-the-lab datasets.

Specifically, EAC-Net, ARL, JÂA-Net and SEV-Net were evaluated on all three datasets. Meanwhile, SRERL, HSTR-Net, MMA-Net, GeoCNN, DC-Net, ME-GraphAU, and WSRTL were evaluated on both BP4D and DISFA datasets. D-PAttNet$^{tt}$ was solely evaluated on BP4D. On the other hand, FACS3D-Net, ML-GCN, and MS-CAM were solely evaluated on BP4D+. However, since BP4D+ was released at a later time, not all methods utilized it.

Following the majority of prior research, we adopted the F1 score on the frame level as the evaluation metric. This metric calculates the harmonic mean of precision (P) and recall (R) and is represented as $F1 = 2PR/(P + R)$. By considering both false positives and false negatives, the F1-score provides a more comprehensive and effective measurement of the model's performance compared to accuracy, especially on datasets where class distribution is uneven. Additionally, some previous studies have reported accuracy and AUC as performance measures. Thus, for comparison purposes, we present the accuracy and AUC on all datasets. It is worth noting that since the F1-score is the most frequently used metric for AU detection, most prior studies only report F1-Score results and rarely provide accuracy and AUC [20, 32, 38, 42].

**Evaluation on BP4D** Table 2 displays the F1-score outcomes of our method on BP4D, while Table 3 shows the accuracy and AUC results. Our approach achieves superior results, with an average F1-score of 66.3%, accuracy of 80.8%, and AUC of 84.2%, surpassing all other state-of-the-art methods.

Table 2 demonstrates that our method achieves an average F1 score of 66.3%, outperforming all other comparison methods by a significant margin. Notably, our

**Table 2** Comparisons of our method and the state-of-the-art methods on **BP4D** in terms of F1 scores (%)

| Method | AU1 | AU2 | AU4 | AU6 | AU7 | AU10 | AU12 | AU14 | AU15 | AU17 | AU23 | AU24 | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| EAC-Net | 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 | 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 |
| SRERL | 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 |
| JÂA-Net | 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 |
| Dual-Learning | 52.6 | 44.9 | 56.2 | 79.8 | [80.4] | [85.2] | 88.3 | 65.6 | 51.7 | 59.4 | 47.3 | 49.2 | 63.4 |
| MMA-Net | 52.5 | [50.9] | 58.3 | 76.3 | 75.7 | 83.8 | 87.9 | 63.8 | 48.7 | 61.7 | 46.5 | [54.4] | 63.4 |
| DC-Net | 51.6 | 47.3 | 56.8 | 79.0 | 79.7 | 84.5 | 88.0 | 65.6 | 51.3 | 62.5 | 47.7 | 51.5 | 63.8 |
| GeoCNN | 48.4 | 44.2 | 59.9 | 78.4 | 75.6 | 83.6 | 86.7 | 65.0 | 53.0 | 64.7 | 49.5 | 54.1 | 63.6 |
| SEV-Net | [58.2] | 50.4 | 58.3 | **[81.9]** | 73.9 | **[87.8]** | 87.5 | 61.6 | 52.6 | 62.2 | 44.6 | 47.6 | 63.9 |
| D-PAttNet$^{tt}$ | 50.7 | 42.5 | 59.0 | 79.4 | 79.0 | 85.0 | 89.3 | [67.6] | 51.6 | **[65.3]** | 49.6 | 54.5 | 64.7 |
| ME-GraphAU | 53.7 | 46.9 | 59.0 | 78.5 | 80.0 | 84.4 | 87.8 | 67.3 | 52.5 | 63.2 | 50.6 | 52.4 | 64.7 |
| HSTR-Net | 55.5 | 49.5 | **[61.9]** | 76.6 | 80.2 | 84.2 | 87.4 | 62.6 | [54.8] | 64.1 | 47.1 | 52.1 | 64.7 |
| WSRTL | **[59.7]** | **[51.7]** | [61.6] | [80.3] | **[80.9]** | [85.2] | [89.7] | **[67.8]** | 52.2 | 63.4 | [51.4] | 46.9 | [65.9] |
| GTLE-Net | [58.2] | 48.7 | 61.5 | 78.7 | 79.2 | 84.2 | **[89.8]** | 66.3 | **[56.7]** | [64.8] | **[53.5]** | 53.6 | **[66.3]** |

The best results are shown in bold and brackets, and the second-best results are in brackets

**Table 3** Comparisons of our method and the state-of-the-art methods on **BP4D** in terms of accuracy (%) and AUC

| AU | Accuracy | | | | | AUC | | | |
|---|---|---|---|---|---|---|---|---|---|
| | EAC | JÂA-Net | ARL | MMA-Net | GTLE-Net | SRERL | Dual-Learning | ME-GraphAU | GTLE-Net |
| 1 | 68.9 | 75.2 | 73.9 | [77.2] | **[81.7]** | 67.6 | [78.5] | 75.0 | **[81.3]** |
| 2 | 73.9 | 80.2 | 76.7 | [82.4] | [80.4] | 70.0 | 75.9 | [78.0] | **[79.8]** |
| 4 | 78.1 | **[82.9]** | [80.9] | [82.9] | **[82.9]** | 73.4 | 84.4 | [85.4] | **[87.0]** |
| 6 | 78.5 | [79.8] | 78.2 | 79.3 | **[80.2]** | 78.4 | [88.6] | [88.9] | **[89.1]** |
| 7 | 69.0 | 72.3 | [74.4] | 73.0 | **[76.2]** | 76.1 | [84.8] | 84.0 | **[84.9]** |
| 10 | 77.6 | 78.2 | 79.1 | [80.2] | **[81.0]** | 80.0 | [87.3] | **[87.4]** | 86.3 |
| 12 | 84.6 | [86.6] | 85.5 | 86.4 | **[88.6]** | 85.9 | [93.9] | 93.2 | **[94.8]** |
| 14 | 60.6 | 65.1 | 62.8 | [66.5] | **[66.0]** | 64.4 | [71.8] | 69.9 | **[72.4]** |
| 15 | 78.1 | 81.0 | [84.7] | 82.9 | **[85.2]** | 75.1 | 80.7 | [82.7] | **[84.0]** |
| 17 | 70.6 | 72.8 | [74.1] | 71.2 | **[76.9]** | 71.7 | 75.0 | [79.1] | **[81.5]** |
| 23 | 81.0 | 82.9 | 82.9 | [83.9] | **[84.5]** | 71.6 | 78.7 | [79.5] | **[82.4]** |
| 24 | 82.4 | [86.3] | 85.7 | **[87.2]** | 85.8 | 74.6 | 84.3 | **[87.8]** | [86.9] |
| Avg | 75.2 | 78.6 | 78.2 | [79.5] | **[80.8]** | 74.1 | 82.0 | [82.6] | **[84.2]** |

The best results are shown in bold and brackets, and the second-best results are in brackets

approach improves upon the second-highest performing supervised-learning-based methods, including D-PAttNe$t^{tt}$, ME-GraphAU, and HSTR-Net, by approximately 1.6%. Moreover, our method surpasses state-of-the-art weakly supervised learning approach WSRTL by approximately 0.4% on BP4D and by a considerable margin on DISFA (refer to "Evaluation on DISFA" for further details). Regarding individual AUs, our method achieves the highest or second-highest F1 scores on 5 out of the 12 evaluated AUs, namely AU1, AU12, AU15, AU17, and AU23. Additionally, our approach exhibits over 1.9% superiority on AU15 and AU23 compared to the second-highest performing methods.

Our method demonstrates superior performance in AU feature learning compared to D-PAttNe$t^{tt}$ by approximately 1.6%. This finding highlights the effectiveness of extracting local features through attention, indicating that our approach outperforms patch learning in AU feature extraction.

The non-predetermined AU attention learning approach exhibits superior performance compared to EAC-Net and JÂA-Net by 10.4% and 3.9%, both of which employ attention mechanisms with pre-defined AU centers as supervision. This underscores the effectiveness of our approach in AU attention learning without the need for predetermined AU centers.

Furthermore, Table 3 demonstrates that our method achieves the highest performance on the average accuracy or AUC, surpassing other methods by more than 1.3% and 1.6%,

**Table 4** Comparisons of our method and the state-of-the-art methods on **DISFA** in terms of F1 scores (%)

| Method | AU1 | AU2 | AU4 | AU6 | AU9 | AU12 | AU25 | AU26 | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| EAC-Net | 41.5 | 26.4 | 66.4 | 50.7 | 8.5 | **[89.3]** | 88.9 | 15.6 | 48.5 |
| ARL | 43.9 | 42.1 | 63.6 | 41.8 | 40.0 | 76.2 | 95.2 | [66.8] | 58.7 |
| SRERL | 45.7 | 47.8 | 56.9 | 47.1 | 45.6 | 73.5 | 84.3 | 43.6 | 55.9 |
| JÂA-Net | 62.4 | 60.7 | 67.1 | 41.1 | 45.1 | 73.5 | 90.9 | 67.4 | 63.5 |
| SEV-Net | 55.3 | 53.1 | 61.5 | 53.6 | 38.2 | 71.6 | [95.7] | 41.5 | 58.8 |
| GeoCNN | **[65.5]** | **[65.8]** | 67.2 | 48.6 | 51.4 | 72.6 | 80.9 | 44.9 | 62.1 |
| DC-Net | 48.7 | 51.6 | 68.2 | [55.6] | 52.3 | 76.0 | 92.6 | 54.3 | 62.4 |
| HSTR-Net | 54.3 | 50.8 | 70.1 | **[66.6]** | [59.6] | 68.0 | **[97.9]** | [69.8] | 62.9 |
| ME-GraphAU | 54.6 | 47.1 | 72.9 | 54.0 | 55.7 | 76.7 | 91.1 | 53.0 | 63.1 |
| Dual-learning | 62.9 | **[65.8]** | 71.3 | 51.4 | 45.9 | 76.0 | 92.1 | 50.2 | 64.4 |
| WSRTL | 57.3 | 51.8 | **[74.3]** | 49.8 | 44.8 | [79.3] | 94.6 | 64.6 | 64.6 |
| MMA-Net | 63.8 | 54.8 | [73.6] | 39.2 | **[61.5]** | 73.1 | [92.3] | **[70.5]** | [66.0] |
| GTLE-Net | [64.5] | [63.2] | 70.1 | 47.7 | 53.6 | 76.2 | 94.8 | 65.1 | **[66.9]** |

The best results are shown in bold and brackets, and the second-best results are in brackets

**Table 5** Comparisons of our method and the state-of-the-art methods on **DISFA** in terms of Accuracy (%) and AUC

| AU | Accuracy | | | | | AUC | | | |
|---|---|---|---|---|---|---|---|---|---|
| | EAC | MMA-Net | JÂA-Net | ARL | GTLE-Net | SRERL | Dual-learning | ME-GraphAU | GTLE-Net |
| 1 | 85.6 | [96.8] | **[97.0]** | 92.1 | 95.4 | 76.2 | **[90.5]** | 90.0 | [90.1] |
| 2 | 84.9 | [96.5] | **[97.3]** | 92.7 | [96.5] | 80.9 | **[92.7]** | 88.5 | [90.2] |
| 4 | 79.1 | **[91.6]** | 88.0 | [88.5] | [90.5] | 79.1 | [93.8] | **[94.2]** | 92.3 |
| 6 | 69.1 | 91.5 | **[92.1]** | [91.6] | 91.5 | 80.4 | [90.3] | **[92.5]** | 90.0 |
| 9 | 88.1 | **[96.5]** | 95.6 | 95.9 | [96.4] | 76.5 | 84.4 | [91.5] | **[92.9]** |
| 12 | 90.0 | 92.3 | 92.3 | **[93.9]** | [93.6] | 87.9 | 95.7 | [95.9] | **[96.4]** |
| 25 | 80.5 | 95.5 | 94.9 | **[97.3]** | [97.1] | 90.9 | [98.2] | **[99.1]** | **[99.1]** |
| 26 | 64.8 | **[95.0]** | [94.8] | [94.8] | 93.9 | 73.4 | 87.4 | **[91.2]** | [91.1] |
| Avg | 80.6 | **[94.5]** | 94.0 | 93.3 | [94.4] | 80.7 | 91.6 | **[92.9]** | [92.8] |

The best results are shown in bold and brackets, and the second-best results are in brackets

respectively. Additionally, our approach yields the highest or second-highest scores for nearly all AUs, regardless of accuracy or AUC.

**Evaluation on DISFA** The F1-score on DISFA is presented in Table 4, while Table 5 displays the accuracy and AUC results on DISFA. Our approach delivers remarkable results, achieving an average F1 score of 66.9%, accuracy of 94.4%, and AUC of 92.8%, surpassing or obtaining comparable outcomes to all compared state-of-the-art methods.

Compared to the most recent works ME-GraphAU, WSRTL and MMA-Net, our method achieves impressive improvements of 3.8%, 2.3% and 0.9%, respectively, on the F1-score. It is worth noting that our model only utilizes AU labels, unlike EAC-Net and JÂA-Net, which are supervised by both AU and landmark labels. Despite this, our approach still outperforms them by 18.4% and 3.4%, respectively, highlighting the effectiveness of our method in AU feature and attention learning. Note that our method does not explicitly model AU relationships, yet it significantly exceeds all relation-based methods, including SRERL and ME-GraphAU. Additionally, Table 5 also provides accuracy and AUC results on DISFA. Our approach achieves comparable accuracy and AUC to those of state-of-the-art methods. Its performance is inferior to the best ones (MMA-Net on Accuracy and ME-GraphAU on AUC) by only 0.1%. It is noteworthy that the F1 score is a better metric [17, 31, 53] (mentioned in Section 4.3) on imbalanced datasets like DISFA, which supports the validation of our approach due to its best performance on the F1 score (See Table 4).

**Evaluation on BP4D+** The evaluation of F1-score on BP4D+ is presented in Table 6. The results reported in [44] include SEV-Net, GCN, and MS-CAM. Additionally, Table 7 lists the accuracy (ACC) and AUC results of our GTLE-Net approach, named GTLE-Net(ACC) and GTLE-Net(AUC), respectively It is noteworthy that those SOTA methods listed in Table 6 do not report metrics of ACC and AUC, nor do they provide their available codes, thus they are not compared in Table 7.

Table 6 demonstrates that our method achieves the highest or second-highest results on 11 out of 12 AUs evaluated, with an average F1 score of 65.7%. Our approach surpasses all the state-of-the-art methods, outperforming the second-best method (SEV-Net) by more than 4.2%. Additionally, Table 7 showcases that our method achieves an impressive accuracy of 88.6% and an AUC of 89.1% on BP4D+.

**Cross-dataset evaluation** JÂA-Net [32] conducts a cross-dataset evaluation to assess the generalization performance on a large-scale test provides. They trained their model on BP4D and evaluated it on the entire BP4D+ dataset comprising 140 subjects. Following their study, we replicate the experiment with our method, and the outcomes are displayed

**Table 6** Comparisons of our method and the state-of-the-art methods on **BP4D+** in terms of F1 scores (%)

| Method | AU1 | AU2 | AU4 | AU6 | AU7 | AU10 | AU12 | AU14 | AU15 | AU17 | AU23 | AU24 | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| FACS3D-Net | 43.0 | 38.1 | **[49.9]** | 82.3 | 85.1 | 87.2 | 87.5 | 66.0 | 48.4 | 47.4 | 50.0 | [31.9] | 59.7 |
| ML-GCN | 40.2 | 36.9 | 32.5 | 84.8 | 88.9 | 89.6 | [89.3] | 81.2 | [53.3] | 43.1 | 55.9 | 28.3 | 60.3 |
| MS-CAM | 38.3 | 37.6 | 25.0 | 85.0 | **[90.9]** | [90.9] | 89.0 | 81.5 | **[60.9]** | 40.6 | 58.2 | 28.0 | 60.5 |
| SEV-Net | [47.9] | **[40.8]** | 31.2 | **[86.9]** | 87.5 | 89.7 | 88.9 | **[82.6]** | 39.9 | **[55.6]** | [59.4] | 27.1 | [61.5] |
| GTLE-Net (F1) | **[51.5]** | [46.6] | [43.5] | [86.8] | [89.6] | **[91.0]** | **[89.8]** | [82.3] | 46.8 | [49.3] | **[60.9]** | **[50.9]** | **[65.7]** |

The best results are shown in bold and brackets, and the second-best results are in brackets

**Table 7** Accuracy(ACC %) and AUC results of our method on **BP4D+**

| Method | AU1 | AU2 | AU4 | AU6 | AU7 | AU10 | AU12 | AU14 | AU15 | AU17 | AU23 | AU24 | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GTLE-Net (ACC) | 90.9 | 90.4 | 94.6 | 87.0 | 86.3 | 88.4 | 88.2 | 77.2 | 89.1 | 87.5 | 87.1 | 96.1 | 88.6 |
| GTLE-Net (AUC) | 87.3 | 84.0 | 83.4 | 94.7 | 92.3 | 95.1 | 94.6 | 82.7 | 87.0 | 85.9 | 87.8 | 94.6 | 89.1 |

The use of these bold entries facilitates readability

in Table 8. We compare our results with those presented in [32]. Once again, our approach exhibits superior performance to all other methods tested under the large-scale cross-dataset evaluation. This indicates that our approach can extract identity-independent features and generalize effectively to new subjects.

**Evaluation on RAF-AU** The results of the evaluation of the test set on RAF-AU are presented in Table 9. Notably, only AU-CNN [42] has been validated with RAF-AU in the literature. To provide further evidence of the efficacy of our approach in the wild, we trained two other state-of-the-art methods, JÂA-Net and ME-GraphAU, using their public codes on RAF-AU. We also conducted an experiment without *GEE*, referred to as *w/o GEE* in Table 9, to assess the contribution of *GEE*.

With respect to F1-score, GTLE-Net records the highest average scores, surpassing ME-GraphAU, AU-CNN and JÂA-Net by 3.75%, 6.82% and 21.4%, respectively. Notably, GTLE-Net also achieves the highest or second-highest F1-score for almost all AUs, indicating its potential in detecting AUs in diverse settings. Additionally, GTLE-Net significantly outperforms other approaches in terms of accuracy and AUC both on average and individual AU results. Our method attains the highest AUC performance across all AUs, surpassing the second highest, AU-CNN, by 2.91% on average AUC. These findings further demonstrate that GTLE-Net exhibits superior performance and robustness even on complex and diverse facial expression data.

Moreover, GTLE-Net outperforms JÂA-Net in F1-score, accuracy and AUC. One contributing factor is the superior generalization ability of the *GEE* module trained on in-the-wild datasets, as demonstrated by the pretraining of GTLE-Net on a facial expression similarity task. The results in Table 9 confirm this, with GTLE-Net achieving improvements of 13.93%, 4.32% and 6.27% in F1-score, accuracy and AUC, respectively, compared to the feature extractor of GTLE-Net being randomly initialized (*w/o GEE*).

In addition, supervised attention-based techniques, such as JÂA-Net, may require landmark detection to determine the location of AUs. However, this step heavily depends on the accuracy of landmark detection and tends to overfit to particular patterns, making it vulnerable to factors like large head pose, varied lighting, and exaggerated facial expressions. This further highlights the superiority and robustness of our approach, which uses adaptive attention.

## 4.4 Ablation study

In this section, we conduct ablation studies to evaluate the effectiveness of each key sub-module in our framework, including: 1) the prior knowledge of expression similarity learning, and 2) the local AU features module. Specifically, all the ablation experiments are conducted on the BP4D dataset due to space limitations. The experimental results are reported in Table 10.

**Partial data results** As mentioned above, GTLE-Net can alleviate the issue of overfitting in training data due to the design of *GEE* and local AU feature learning. This means that it can perform well even with limited training data. To confirm this, we trained GTLE-Net, JÂA-Net, and MeGraph-AU on only 10% and 1% of the available training data, respectively. The results of F1-score and accuracy are presented in Fig. 4. While all three methods experienced a decline in performance with reduced training data, GTLE-Net demonstrated a slower decline compared to the other two methods. It maintained the highest F1-score, particularly on the in-the-wild dataset, despite having extremely limited training data.

**Table 8** Comparisons of our method and the state-of-the-art on cross-dataset evaluation (**trained on BP4D and evaluated on BP4D+**) in terms of F1 scores (%)

| Method | AU1 | AU2 | AU4 | AU6 | AU7 | AU10 | AU12 | AU14 | AU15 | AU17 | AU23 | AU24 | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| EAC-Net | 38.0 | [37.5] | [32.6] | 82.0 | 83.4 | 87.1 | 85.1 | 62.1 | 44.5 | 43.6 | 45.0 | 32.8 | 56.1 |
| ARL | 29.9 | 33.1 | 27.1 | 81.5 | 83.0 | 84.8 | 86.2 | 59.7 | [44.6] | 43.7 | 48.8 | 32.3 | 54.6 |
| JÂA-Net | **[39.7]** | 35.6 | 30.7 | [82.4] | [84.7] | [88.8] | [87.0] | [62.2] | 38.9 | **[46.4]** | [48.9] | **[36.6]** | [56.8] |
| GTLE-Net | [39.5] | **[37.7]** | **[44.0]** | **[84.5]** | **[84.9]** | **[89.6]** | **[87.7]** | **[72.3]** | **[44.9]** | [45.3] | **[52.6]** | [33.9] | **[59.7]** |

The best results are shown in bold and brackets, and the second-best results are in brackets

**Table 9** Comparisons of our method and the state-of-the-art methods on **RAF-AU** in terms of F1 scores (%), Accuracy(%) and AUC

| | F1 | | | | | Accuracy | | | | AUC | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AU | AU-CNN | JÂA-Net | ME-GraphAU | w/o GEE | GTLE-Net | JÂA-Net | ME-GraphAU | w/o GEE | GTLE-Net | AU-CNN | JÂA-Net | ME-GraphAU | w/o GEE | GTLE-Net |
| 1 | 60.47 | 38.77 | [69.07] | 45.23 | **[71.46]** | 76.65 | [84.23] | 76.87 | **[86.77]** | [82.03] | 54.53 | 79.47 | 75.46 | **[89.44]** |
| 2 | 65.59 | 52.63 | [69.59] | 51.77 | **[72.27]** | 85.99 | [89.64] | 84.35 | **[89.76]** | **[90.49]** | 55.59 | 80.27 | 82.42 | [90.31] |
| 4 | 73.44 | 65.03 | **[81.15]** | 71.04 | [81.14] | 69.45 | [84.93] | 76.64 | **[85.50]** | [86.28] | 67.91 | 83.87 | 83.88 | **[92.61]** |
| 5 | 69.69 | 38.99 | [68.32] | 55.62 | **[74.79]** | 81.13 | [88.26] | 82.74 | **[89.53]** | [88.86] | 53.57 | 77.02 | 84.93 | **[91.74]** |
| 6 | [58.21] | 34.29 | 53.80 | 40.56 | **[60.34]** | 91.05 | [90.91] | 90.22 | **[91.83]** | [87.41] | 54.55 | 72.47 | 81.87 | **[87.99]** |
| 9 | 67.44 | 40.00 | **[72.73]** | 55.51 | [71.52] | 83.07 | **[92.06]** | 87.46 | [90.10] | [90.03] | 56.69 | 81.76 | 84.49 | **[92.35]** |
| 10 | 68.41 | 52.01 | [70.86] | 63.89 | **[74.23]** | 74.51 | [80.78] | 78.02 | **[82.74]** | [87.89] | 61.24 | 79.32 | 83.10 | **[90.80]** |
| 12 | 69.62 | 64.41 | 67.37 | [71.26] | **[73.02]** | [83.66] | [83.66] | 83.31 | **[84.70]** | [88.48] | 69.96 | 80.64 | 86.63 | **[90.58]** |
| 16 | **[59.38]** | 45.45 | 52.40 | 53.28 | [55.45] | 85.99 | **[87.46]** | [86.88] | 84.47 | [86.11] | 63.54 | 69.91 | 86.76 | **[88.94]** |
| 17 | 25.64 | 29.55 | [62.75] | 59.48 | **[64.76]** | 87.94 | [89.18] | 85.96 | **[90.10]** | [88.58] | 62.70 | 76.24 | 85.54 | **[91.27]** |
| 25 | [92.14] | 89.72 | 91.84 | 90.91 | **[94.89]** | 86.58 | [88.15] | 86.31 | **[93.67]** | [95.33] | 88.59 | 86.24 | 93.63 | **[97.97]** |
| 26 | [64.65] | 44.76 | 63.57 | 53.80 | **[65.08]** | 77.43 | [81.82] | 80.90 | **[83.08]** | [86.31] | 58.60 | 69.73 | 85.15 | **[88.87]** |
| 27 | [84.67] | 73.47 | 82.95 | 85.21 | **[88.42]** | 92.41 | [94.25] | 92.06 | **[95.63]** | 95.77 | 78.73 | 91.98 | [95.94] | **[98.47]** |
| Avg | 65.95 | 51.47 | [69.12] | 58.94 | **[72.87]** | 82.76 | [87.33] | 83.98 | **[88.30]** | [88.73] | 63.55 | 79.15 | 85.37 | **[91.64]** |

The best results are shown in bold and brackets, and the second-best results are in brackets

**Table 10** Ablation study on BP4D measured by F1 scores (%)

| Method | AU1 | AU2 | AU4 | AU6 | AU7 | AU10 | AU12 | AU14 | AU15 | AU17 | AU23 | AU24 | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *w/o pretrain* | 34.9 | 24.2 | 36.7 | 76.1 | 71.4 | 81.0 | 86.3 | 58.7 | 25.9 | 51.9 | 27.4 | 31.6 | 50.5 |
| *w. GEE fixed* | 52.3 | 45.8 | 59.8 | **[79.4]** | 78.1 | **[84.8]** | 89.0 | 64.0 | 46.5 | 64.0 | 38.6 | 49.8 | 63.5 |
| *w. GEE Emotion* | 53.7 | 44.4 | 57.9 | 78.9 | 78.9 | **[84.8]** | 89.2 | [65.8] | 50.0 | 63.5 | 50.5 | 53.4 | 64.3 |
| *w. GEE ImageNet* | 49.6 | 39.7 | 60.5 | 77.3 | 78.7 | 83.1 | 88.6 | 64.6 | 48.2 | 63.2 | 47.5 | 48.0 | 62.4 |
| *w. VGG16 ImageNet* | 50.6 | 44.9 | 53.9 | 76.5 | [79.3] | 83.4 | 89.0 | **[66.3]** | 47.6 | 61.7 | 46.4 | 45.0 | 62.1 |
| *w. RepVGG-A2 ImageNet* | 54.6 | 42.1 | 55.8 | 78.1 | 77.7 | 84.1 | 87.4 | 65.1 | 49.5 | 63.6 | 46.1 | 45.7 | 62.5 |
| *w. ResNet50 ImageNet* | [56.4] | 46.5 | 57.1 | 78.1 | 79.0 | 84.2 | 87.3 | 63.3 | 51.0 | 64.3 | 49.2 | 53.2 | 64.1 |
| *w/o* $E_m$&$E_a$ | 55.4 | **[65.1]** | **[73.9]** | 49.0 | 46.9 | 77.1 | **[92.9]** | 54.1 | 48.1 | 64.7 | [52.5] | 47.5 | 63.3 |
| *w/o* $E_m$ | 51.9 | 43.0 | 59.2 | **[79.4]** | 78.2 | [84.5] | 89.1 | 63.2 | 50.6 | 64.4 | 49.3 | **[55.2]** | 64.0 |
| *w. Supervised-attention* | 54.3 | 45.9 | 61.3 | [79.3] | **[79.4]** | 84.3 | 89.2 | 62.1 | [51.6] | **[65.5]** | 50.1 | 53.0 | [64.7] |
| GTLE-Net (Full) | **[58.2]** | [48.7] | [61.5] | 78.7 | 79.2 | 84.2 | [89.8] | **[66.3]** | **[56.7]** | [64.8] | **[53.5]** | [53.6] | **[66.3]** |

The best results are shown in bold and brackets, and the second-best results are in brackets.

These findings support our earlier assertion that GTLE-Net is more robust and has better overall generalization capability.

**Prior knowledge** The proposed framework introduces a feature encoder, namely $GEE$, which is pre-trained on the FEC dataset using expression similarity learning based on our prior knowledge. To evaluate the effectiveness of this prior knowledge, we conducted ablation experiments to initialize $GEE$ using various approaches during the training of *GTLE-Net* as follows:

a) *w/o pretrain*: $GEE$ is randomly initialized;
b) *w. GEE Emotion*: $GEE$ is initialized via a pretrained classifier of seven discrete facial expression emotion (e.g. neutral, happy, sad, surprise, fear, disgust, angry) categories trained on the AffectNet dataset;
c) *w. GEE ImageNet*: $GEE$ is initialized via a pretrained image object classifier built on imageNet [14];
d) *GTLE-Net (Full)*: our proposed method, $GEE$ is initialized via the facial expression similarity learning;

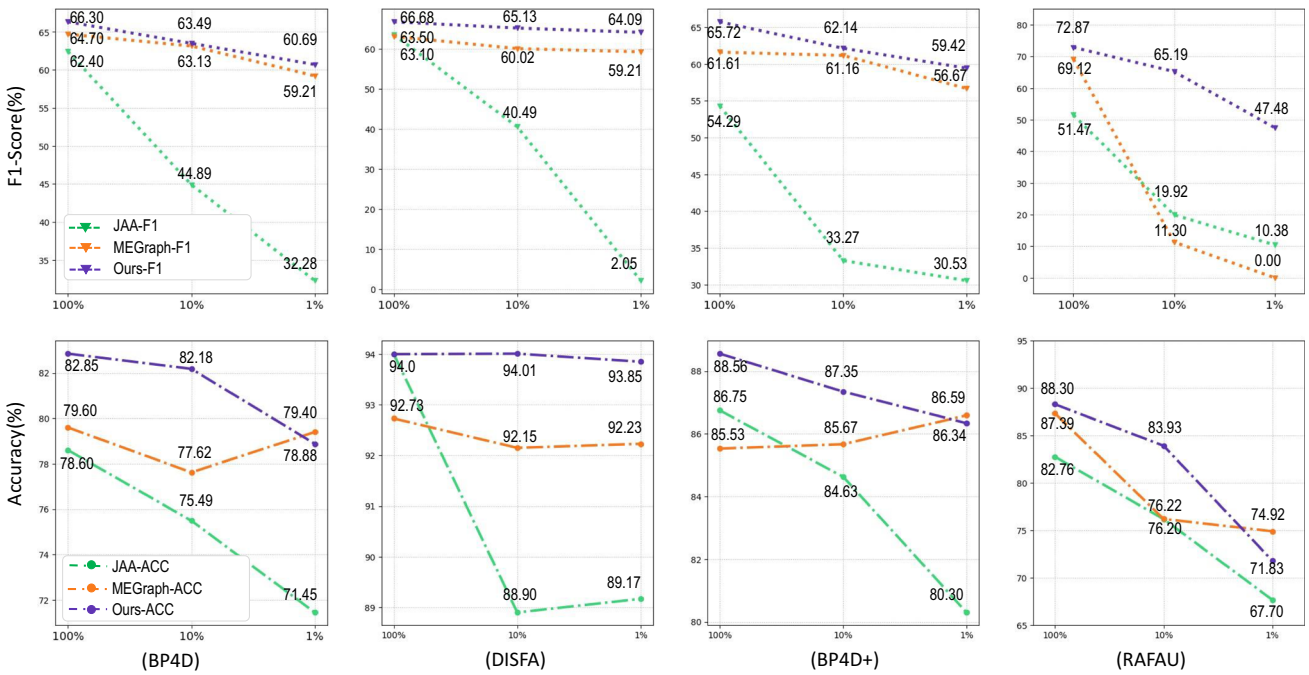



**Fig. 4** Visualization of the partial data (100%, 10% and 1%) ablation experiments(zoom in for details). The top row is F1-Score (%) and the bottom row is accuracy (%). Each row includes 4 columns of the partial data experiments (BP4D, BP4D+, DISFA and RAF-AU)

e) *w. GEE fixed*: $GEE$ is parameterized by the facial expression similarity approach but not updated in the training of *GTLE-Net*.
f) *w. VGG16 ImageNet*: $GEE$ is replaced by VGG16 that is pretrained on ImageNet.
g) *w. RepVGG-A2 ImageNet*: $GEE$ is replaced by RepVGG-A2 that is pretrained on ImageNet.
h) *w. ResNet50 ImageNet*: $GEE$ is replaced by ResNet50 that is pretrained on ImageNet.

Firstly, the results outlined in Table 10 demonstrate that *w/o pretrain* is outperformed by *w. GEE fixed*, *w.GEE Emotion*, and *w. GEE ImageNet* approaches, achieving enhancements of 13% (from 50.5% to 63.5%), 13.8% (from 50.5% to 64.3%), and 11.9% (from 50.5% to 62.4%), respectively. This indicates that prior knowledge is essential, and the pre-training phase can significantly enhance AU detection.

Secondly, when compared to *w. GEE Emotion*, *w. GEE ImageNet*, *w. VGG16 ImageNet*, *w. RepVGG-A2 ImageNet*, and *w. ResNet50 ImageNet*, the performance of *GTLE-Net* (66.3%) is superior, surpassing the other pretraining methods by 2%, 3.9%, 4.2%, 3.8%, and 2.2%, respectively. This highlights the effectiveness and superiority of the prior knowledge captured by our facial expression similarity learning.

Thirdly, in comparison to *GTLE-Net*, *w. GEE fixed* achieves a lower average F1-score of 63.5%. This suggests that fine-tuning the global expression feature encoder is advantageous in extracting more adaptable features for the subsequent *LAM* module and is crucial and effective to enhance the performance.

**Local AU features module** We perform three ablation experiments to verify the effectiveness of the representations acquired by our local AU features module (*LAM*). Specifically, we modify our complete *GTLE-Net* framework by removing or replacing certain sub-modules and then retrain the entire network. The ablation experiments are:

i) *w/o* $E_m$&$E_a$: removing both the $E_m$ and the $E_a$;
j) *w/o* $E_m$: only removing $E_m$;
k) *w. Supervised-attention*: adding the attention map supervision based on landmarks following [32]. To further explore the effect of AU attention based on landmarks during training.

It is worth noting that incorporating $E_a$ alone (*w/o* $E_m$) resulted in an improvement of 0.7% (from 63.3% to 64.0%) when compared to the case of *w/o* $E_m$&$E_a$. Furthermore, with the addition of the $E_m$ module (*GTLE-Net*), the F1-Score increased by 2.3%, thereby validating the effectiveness of our AU mask and local feature learning. In addition, the training and validation losses plot in Fig. 5 reveal that, despite having higher training loss, GTLE-Net produced lower validation loss values when compared to *w/o* $E_m$. This signifies that the local AU features learning ($E_m$) effectively alleviates overfitting.

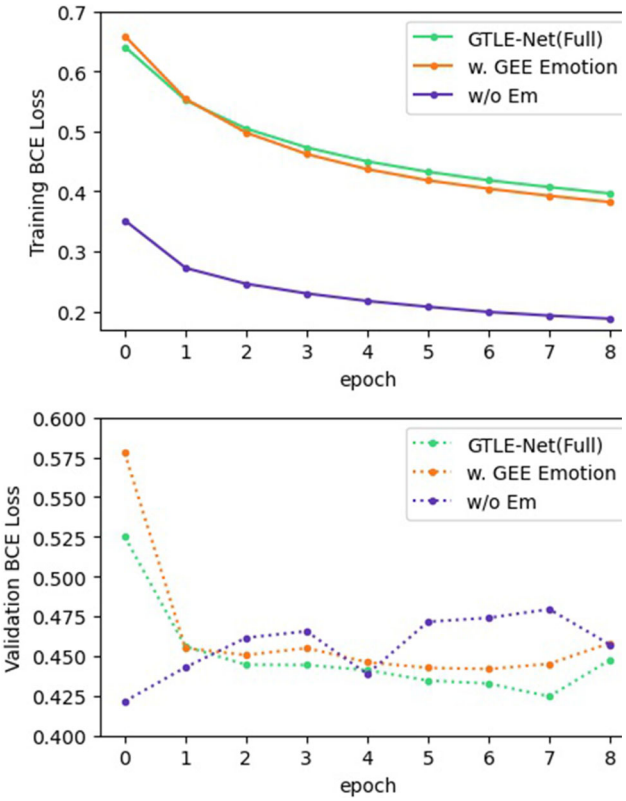



**Fig. 5** Training (up) and validation (down) loss for *GTLE-Net*, *w. GEE Emotion* and *w/o* $E_m$ on the third fold of BP4D dataset

In addition, the F1-score of *w. Supervised-attention* reduced by 1.6% upon the addition of landmark-based AU attention map supervision on the $E_m$. This suggests that our self-adaptive attention map learning is more effective, while the performance of AU detection is affected by the suboptimal performance of landmark detection in *w. Supervised-attention*.

## 4.5 Visual analysis

To further substantiate the efficacy of our framework, we visualize the features and mask results obtained from it.

**GEE feature** Firstly, Fig. 6 displays the global expression representations generated by *GEE* using BP4D samples, where each color represents a different kind of AU combination (i.e., similar expressions) from various identities. Notably, the expression embeddings of samples with similar expressions but different identities are well-clustered, indicating that our *GEE* is effective in producing identity-invariant expression representations. Moreover, to validate the effec-

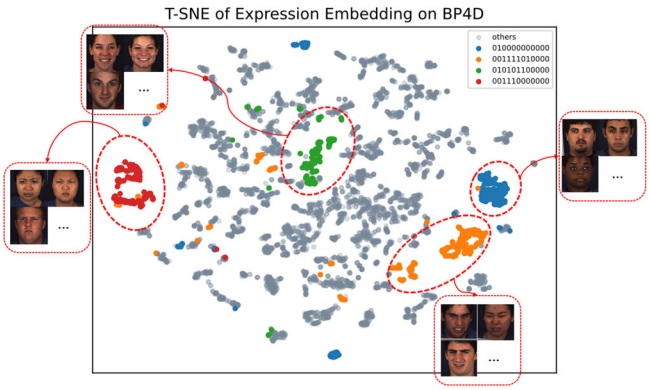



**Fig. 6** Examples of GEE representation distribution on the BP4D dataset

tiveness of disentanglement from other factors like head pose, background, hair, and so on, we conduct image retrieval on the in-the-wild RAF-AU dataset. The retrieval results illustrated in Fig. 7 show that GEE can effectively disentangle expression from other factors. Additionally, the training and validation losses depicted in Fig. 5 demonstrate that *GTLE-Net(Full)* with *GEE* outperforms *w. GEE Emotion* in terms of lower validation loss, indicating that *GEE* can help mitigate the issue of over-fitting.

**Mask attention learning** As per its anatomical definition, each AU is associated with a particular region on the face and requires the activation of a group of muscles. For instance, AU1 involves inner brow raising, while AU9 involves nose wrinkling. Therefore, theoretically, our framework should be able to learn a regional mask activation map for each AU, which corresponds to its respective muscle actions.

The attention mask maps learned by our network for various examples from BP4D, RAF-AU, DISFA, and BP4D+ are illustrated in Figs. 8, 9 and 10. The figure showcases that our method is able to capture distinct attention masks for the different AUs. Specifically, for AU4 (brow low) and AU17 (chin raise), the learned spatial masks were concentrated around the brows and chin respectively, indicating that our suggested AU mask extractor is adequate for learning AU-related local regional attention.

The muscle regions associated with each AU are closely related, as depicted in Fig. 8, even for various expressions such as happy, sad, and fear. It is noteworthy that some AUs may overlap in their coverage of the same or similar areas. This is due to the similarity in the muscle distributions of different AUs. For instance, the forehead is involved in AUs 1, 2, and 4, while the mouth area is closely linked to AUs 14, 15, 17, 23, and 24. Additionally, the masks learned through AU labeling supervision may include other facial regions of correlated AUs due to the potential mutual relations among different AUs. For example, the masks of AUs 1, 2, and 4 not only cover the brow regions but also include the mouth area. Additionally, Fig. 9 demonstrates that the learned AU-related mask maps are also evident on the in-the-wild dataset RAF-AU.

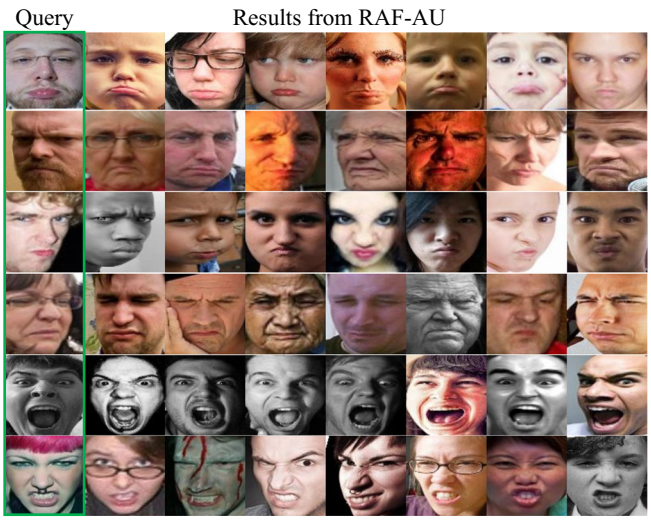



**Fig. 7** Examples of image retrieval on the in-the-wild RAF-AU dataset. This figure shows the top-7 images retrieved using GEE representation. Observing each row, the facial expressions of the retrieved results are closely similar to the query images framed with the green box. As observed, the GEE representation is largely disentangled from other factors like head pose, background, hair, and so on

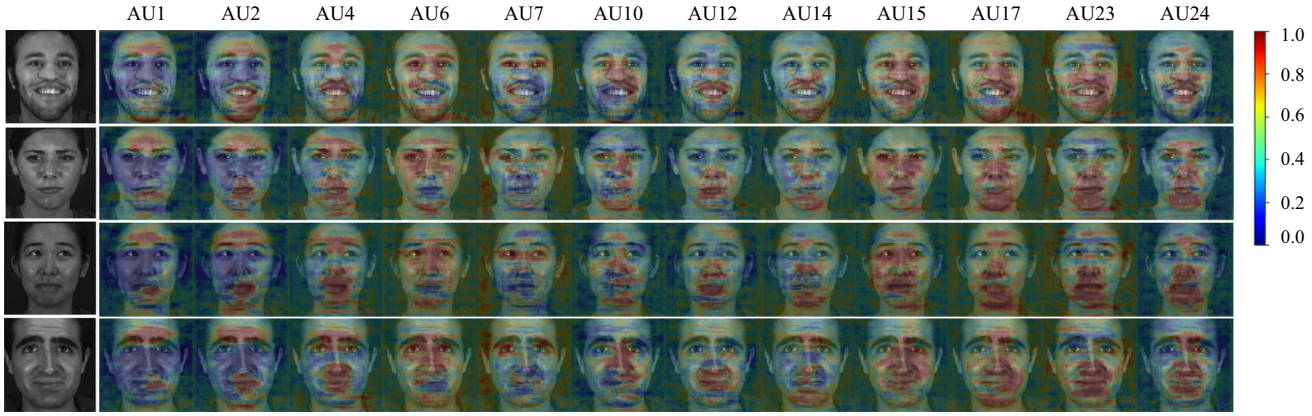



**Fig. 8** Visualization of the mask attention maps overlaid onto the corresponding input images in BP4D. The first column shows face images with different expressions. We can observe that similar AU-related local mask distributions can be learned by our framework even though the identities and expressions vary significantly

## 5 Discussion and conclusion

The proposed framework in this paper involves a unique approach to detect AUs. It involves two learning components, global fine-grained facial expression representation, and local facial attention. The goal of expression representation learning is to obtain strong and reliable latent features that can detect subtle transformations in expressions. Meanwhile, a local AU feature module (*LAM*) is created to generate local attention maps that highlight spatial significance and extract local AU representations.

The proposed framework was extensively tested against state-of-the-art methods on widely-used benchmark datasets (BP4D, DISFA, BP4D+, and RAF-AU) for the AU detection task. The results of our experiments demonstrate that our framework outperforms all other methods significantly. The ablation studies confirm the effectiveness of both fine-grained global expression representation and the *LAM*. Moreover, we demonstrate the robustness and generalization capability of our method through cross-dataset validation, partial dataset experiments, and analysis of the training and validation losses.

Although our method has demonstrated superior performance on four widely-used datasets, there is still room for improvement in detecting AUs. Additionally, due to the limitations of AU annotation, our method can only detect a subset of AUs in FACS. Moreover, our method does not take into account the correlation between AUs, which may affect performance. Although our method is primarily designed for AU detection, it has the potential to be applied to other facial analysis tasks, such as expression recognition and micro-expression detection. In the future, we plan to explore these research directions and extend our current framework.


**Fig. 9** Visualization of the mask attention maps overlaid onto the corresponding input images in RAF-AU

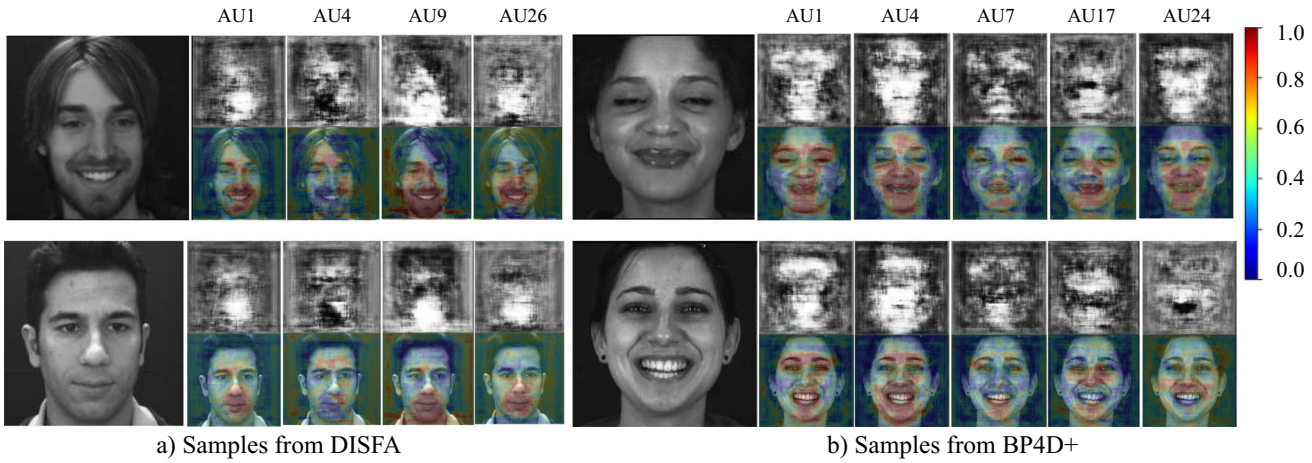



**Fig. 10** Subfigures a) and b) are visualization of the produced mask attention maps: a) samples from DISFA, and b) samples from BP4D+. The learned masks in (odd rows) are also visualized and overlaid on the corresponding face images (even rows). The colors covering the face images from blue to red indicate the attention values from low to high

**Author Contributions** Conceptualization[Rudong An], [Yu Ding], [Wei Zhang], [Hao Zeng], [Zhigang Deng], [Aobo Jin]; Methodology: [Rudong An], [Wei Zhang], [Hao Zeng], [Yu Ding], [Wei Chen]; Investigation: [Rudong An]; Data curation: [Aobo Jin], [Wei Chen]; Writing-review and editing: [Rudong An],[Wei Zhang], [Hao Zeng], [Yu Ding], [Zhigang Deng], [Aobo Jin], [Wei Chen];

**Funding** This work is supported by the 2022 Hangzhou Key Science and Technology Innovation Program (No. 2022AIZD0054) and the Key Research and Development Program of Zhejiang Province (No. 2022C01011).

**Data Availability** All data used in the paper are released and available, including BP4D [49], DISFA [22], BP4D+ [50] and RAF-AU [42].

## Declarations

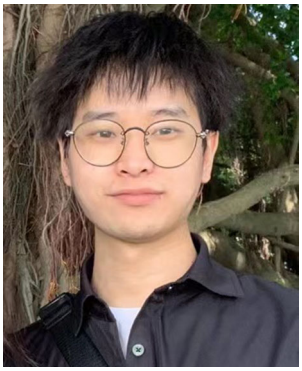
**Competing Interests** All authors declare that they have no conflicts of interest.

**Ethics approval:** This article does not contain any studies with human participants or animals performed by any of the authors.

## References

1. Chen J, Wang C, Wang K et al (2022) Lightweight network architecture using difference saliency maps for facial action unit detection. App Intell 1–22
2. Chen Y, Song G, Shao Z et al (2022) Geoconv: geodesic guided convolution for facial action unit recognition. Pattern Recogn 122:108–355
3. Chen ZM, Wei XS, Wang P et al (2019) Multi-label image recognition with graph convolutional networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5177–5186
4. Choi Y, Uh Y, Yoo J et al (2020) Stargan v2: diverse image synthesis for multiple domains. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8188–8197
5. Cui Z, Song T, Wang Y et al (2020) Knowledge augmented deep neural networks for joint facial expression and action unit recognition. Adv Neural Inf Process Syst 33
6. Ekman P, Friesen W (1978) Facial action coding system: a technique for the measurement of facial movement. Consulting Psychologists Press Palo Alto 12
7. Ertugrul IÖ, Jeni LA, Cohn JF (2019) Pattnet: patch-attentive deep network for action unit detection. In: BMVC, p 114
8. Geng Z, Cao C, Tulyakov S (2019) 3d guided fine-grained face manipulation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9821–9830
9. He K, Zhang X, Ren S et al (2015) Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision, pp 1026–1034
10. He K, Zhang X, Ren S et al (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
11. Hu X, Zhi R, Zhou C (2023) Drop-relationship learning for semi-supervised facial action unit recognition. Neurocomputing p 126361
12. Huang X, Belongie S (2017) Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE international conference on computer vision, pp 1501–1510
13. Jacob GM, Stenger B (2021) Facial action unit detection with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7680–7689
14. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 25
15. Li G, Zhu X, Zeng Y et al (2019) Semantic relationships guided representation learning for facial action unit recognition. In: Proceedings of the AAAI conference on artificial intelligence, pp 8594–8601
16. Li L, Wang S, Zhang Z et al (2021) Write-a-speaker: text-based emotional and rhythmic talking-head generation. In: Proceedings of the AAAI conference on artificial intelligence, pp 1911–1920
17. Li W, Abtahi F, Zhu Z et al (2018) Eac-net: deep nets with enhancing and cropping for facial action unit detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 40(11):2583–2596
18. Liu M, Li S, Shan S et al (2015) Au-inspired deep networks for facial expression feature learning. Neurocomputing 159:126–136

19. Liu S, Wang H (2023) Talking face generation via facial anatomy. ACM Trans Multimedia Comput Commun Appl 19(3)
20. Luo C, Song S, Xie W et al (2022) Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition. In: Raedt LD (ed) Proceedings of international joint conference on artificial intelligence, pp 1239–1246
21. Ma C, Chen L, Yong J (2019) Au r-cnn: encoding expert prior knowledge into r-cnn for action unit detection. Neurocomputing 355:35–47
22. Mavadati SM, Mahoor MH, Bartlett K et al (2013) Disfa: a spontaneous facial action intensity database. IEEE Trans Affect Comput 4(2):151–160
23. Mollahosseini A, Hasani B, Mahoor MH (2017) Affectnet: a database for facial expression, valence, and arousal computing in the wild. IEEE Trans Affect Comput 10(1):18–31
24. Niu X, Han H, Yang S et al (2019) Local relationship learning with person-specific shape regularization for facial action unit detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 11,917–11,926
25. Onal Ertugrul I, Yang L, Jeni LA et al (2019) D-pattnet: dynamic patch-attentive deep network for action unit detection. Frontiers in Computer Science 1:11
26. Pantic M, Rothkrantz L (2004) Facial action recognition for facial expression analysis from static face images. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 34:1449–1461
27. Paysan P, Knothe R, Amberg B et al (2009) A 3d face model for pose and illumination invariant face recognition. In: IEEE international conference on advanced video and signal based surveillance, pp 296–301
28. Rubinow DR, Post RM (1992) Impaired recognition of affect in facial expression in depressed patients. Biological psychiatry 31(9):947–953
29. Shang Z, Du C, Li B et al (2023) Mma-net: multi-view mixed attention mechanism for facial action unit detection. Pattern Recognition Letters
30. Shao Z, Liu Z, Cai J et al (2018) Deep adaptive attention for joint facial action unit detection and face alignment. In: Proceedings of the European conference on computer vision (ECCV), pp 705–720
31. Shao Z, Liu Z, Cai J et al (2019) Facial action unit detection using attention and relation learning. IEEE Transactions on Affective Computing
32. Shao Z, Liu Z, Cai J et al (2021) Jaa-net: joint facial action unit detection and face alignment via adaptive attention. International Journal of Computer Vision 129(2):321–340
33. Song W, Shi S, Dong Y et al (2022) Heterogeneous spatio-temporal relation learning network for facial action unit detection. Pattern Recognition Letters 164:268–275
34. Szegedy C, Ioffe S, Vanhoucke V et al (2017) Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-first AAAI conference on artificial intelligence
35. Ulyanov D, Vedaldi A, Lempitsky V (2016) Instance normalization: The missing ingredient for fast stylization. arXiv:1607.08022
36. Vemulapalli R, Agarwala A (2019) A compact embedding for facial expression similarity. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5683–5692
37. Wang S, Peng G (2019) Weakly supervised dual learning for facial action unit recognition. IEEE Transactions on Multimedia 21(12):3218–3230
38. Wang S, Chang Y, Wang C (2021) Dual learning for joint facial landmark detection and action unit recognition. IEEE Transactions on Affective Computing
39. Xiang X, Tran TD (2017) Linear disentangled representation learning for facial actions. IEEE Transactions on Circuits and Systems for Video Technology 28(12):3539–3544
40. Yan J, Wang J, Li Q et al (2022) Weakly supervised regional and temporal learning for facial action unit recognition. IEEE Transactions on Multimedia
41. Yan J, Wang J, Li Q et al (2022) Weakly supervised regional and temporal learning for facial action unit recognition. IEEE Transactions on Multimedia pp 1–1
42. Yan W, Li S, Que C et al (2020) Raf-au database: in-the-wild facial expressions with subjective emotion judgement and objective au annotations. In: Proceedings of the Asian Conference on Computer Vision (ACCV)
43. Yang B, Wu J, Ikeda K et al (2023) Deep learning pipeline for spotting macro-and micro-expressions in long video sequences based on action units and optical flow. Pattern Recogn Lett 165:63–74
44. Yang H, Yin L, Zhou Y et al (2021) Exploiting semantic embedding and visual feature for facial action unit detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10,482–10,491
45. Yang L, Ertugrul IO, Cohn JF et al (2019) Facs3d-net: 3d convolution based spatiotemporal representation for action unit detection. In: 2019 8th International conference on affective computing and intelligent interaction (ACII), pp 538–544
46. Yao G, Yuan Y, Shao T et al (2021) One-shot face reenactment using appearance adaptive normalization. In: Proceedings of the AAAI conference on artificial intelligence, pp 3172–3180
47. You R, Guo Z, Cui L et al (2020) Cross-modality attention with semantic graph embedding for multi-label classification. In: Proceedings of the AAAI conference on artificial intelligence, pp 12,709–12,716
48. Zhang W, Ji X, Chen K et al (2021) Learning a facial expression embedding disentangled from identity. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 6759–6768
49. Zhang X, Yin L, Cohn JF et al (2014) Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. Image and Vision Computing 32(10):692–706
50. Zhang Z, Girard JM, Wu Y et al (2016) Multimodal spontaneous emotion corpus for human behavior analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3438–3446
51. Zhao K, Chu WS, De la Torre F et al (2015) Joint patch and multi-label learning for facial action unit detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2207–2216
52. Zhao K, Chu WS, Martinez AM (2018) Learning facial action units from web images with scalable weakly supervised clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2090–2099
53. Zhi R, Liu M, Zhang D (2020) A comprehensive survey on automatic facial action unit analysis. The Visual Computer 36(5):1067–1093
54. Zhong L, Liu Q, Yang P et al (2015) Learning multiscale active facial patches for expression analysis. IEEE Transactions on Cybernetics 45(8):1499–1510



**Rudong An** received the B.S. degree and M.S. degree in college of Biomedical Engineering and Instrument Science from Zhejiang University, Hangzhou, China, in 2018 and 2021. He is currently an artificial intelligence researcher at Netease Fuxi AI Lab. His interests include deep learning, action unit detection, facial expression analysis and facial animation generation.

**Aobo Jin** received the BS degree in electrical engineering from the Dalian University of Technology, Dalian, China, in 2011, and the MSc degree in electrical engineering from the Department of Electrical and Computer Engineering, University of Houston, Houston, Texas, in 2016. He is currently working toward the PhD degree with the Department of Computer Science, University of Houston, Houston, Texas. His research interests include computer graphics, virtual human modeling and animation, and sketch-based modeling.

**Wei Chen** received the B.S. degree in information engineering from the National University of Defense Technology, Changsha, China, and received M.S. degree in software engineering from Hebei University, Baoding, China. He is currently an engineer of Hebei Agricultural University, Baoding, China. His interests include software engineering, intelligent optimization algorithm, and deep learning.

**Wei Zhang** received the B.E. degree in communication engineering from Nanjing University of Posts and Telecommunications, Jiangsu, China in 2017 and M.S. degree in electronic and information engineering from Zhejiang University, Zhejiang, China in 2020. She is currently a research scientist working with Netease Fuxi AI Lab, Hangzhou, China. Her current research interests include computer vision, expression embedding and facial affective analysis.

**Hao Zeng** received the B.E. degree in computer software engineering and the M.S. degree in computer science and technology from Huazhong University of Science and Technology, Hubei, China, in 2017 and 2020, respectively. He is currently a researcher in Netease Fuxi AI Lab, Hangzhou, China. His current research interests include deep learning, computer vision and face generation.

**Zhigang Deng** is Moores Professor of Computer Science at University of Houston, Texas, USA. His research interests include computer graphics, computer animation, virtual humans, human computer conversation, and robotics. He earned his Ph.D. in Computer Science at the Department of Computer Science at the University of Southern California in 2006. Prior that, he also completed B.S. degree in Mathematics from Xiamen University (China), and M.S. in Computer Science from Peking University (China). Besides serving as the conference/program co-chair for CASA 2014, SCA 2015, and MIG 2022, he has been an Associate Editor for IEEE Transactions on Visualization and Computer Graphics, Computer Graphics Forum, Computer Animation and Virtual Worlds Journal, etc. He is a distinguished member of ACM and a senior member of IEEE.

**Yu Ding** is currently an artificial intelligence expert, leading the virtual human group at Netease Fuxi AI Lab, Hangzhou, China. His research interests include deep learning, image and video processing, talking face generation, animation generation, facial expression recognition, multimodal computing, affective computing, nonverbal communication (face, gaze, and gesture), and embodied conversational agent. He received Ph.D. degree in Computer Science (2014) at Telecom Paristech in Paris (France), M.S. degree in Computer Science at Pierre and Marie Curie University (France), and B.S. degree in Automation at Xiamen University (China).

## Authors and Affiliations

**Rudong An[1] · Aobo Jin[2] · Wei Chen[3] · Wei Zhang[1] · Hao Zeng[1] · Zhigang Deng[4] · Yu Ding[1]**

Rudong An
anrudong@corp.netease.com

Aobo Jin
jina@uhv.edu

Wei Chen
rshchchw@hebau.edu.cn

Wei Zhang
zhangwei05@corp.netease.com

Hao Zeng
zenghao03@corp.netease.com

Zhigang Deng
zdeng4@central.uh.edu

[1] Virtual Human Group, Netease Fuxi AI Lab, Hangzhou, China

[2] University of Houston-Victoria, Houston, USA

[3] Hebei Agricultural University, Hebei, China

[4] University of Houston, Houston, TX, USA