Detecting Facial Action Units from Global-Local Fine-grained Expressions

Wei Zhang, Lincheng Li, Yu Ding*, Wei Chen, Zhigang Deng, Xin Yu

***Abstract*—Since Facial Action Unit (AU) annotations require domain expertise, common AU datasets only contain a limited number of subjects. As a result, a crucial challenge for AU detection is addressing identity overfitting. We find that AUs and facial expressions are highly associated, and existing facial expression datasets often contain a large number of identities. In this paper, we aim to utilize the expression datasets without AU labels to facilitate AU detection. Specifically, we develop a novel AU detection framework aided by the Global-Local facial Expressions Embedding, dubbed GLEE-Net. Our GLEE-Net consists of three branches to extract identity-independent expression features for AU detection. We introduce a global branch for modeling the overall facial expression while eliminating the impacts of identities. We also design a local branch focusing on specific local face regions. The combined output of global and local branches is firstly pre-trained on an expression dataset as an identity-independent expression embedding, and then finetuned on AU datasets. Therefore, we significantly alleviate the issue of limited identities. Furthermore, we introduce a 3D global branch that extracts expression coefficients through 3D face reconstruction to consolidate 2D AU descriptions. Finally, a Transformer-based multi-label classifier is employed to fuse all the representations for AU detection. Extensive experiments demonstrate that our method significantly outperforms the state-of-the-art on the widely-used DISFA, BP4D and BP4D+ datasets.**

***Index Terms*—Action Units; facial expression; expression embedding; deep learning;**

I. Introduction

The facial expression analysis has been an important challenge. Recently, researchers make considerable efforts to expression representation [1], [59], expression categories [2], [3], expression recognition [4]–[7], expression synthesis [8]–[17], [55], and multimodal sentiment analysis [18]–[25]. As a fundamental issue of these research topics, the detection of facial action units has not been well studied.

Facial Action Units (AUs), coded by Facial Action Coding System (FACS), are defined to describe the local movement of facial expressions based on facial muscle groups [26]. For instance, as shown in Fig. 1 , AU4 represents the movement of lowering brow, which often occurs when expressing anger. AU12 stands for the lip corner puller, which often appears with a happy expression. Face AU detection has attracted lots of research efforts due to its crucial applications in emotion recognition [27], micro-expression detection [28], face synthesis [29]–[31], and mental health diagnosis [32].

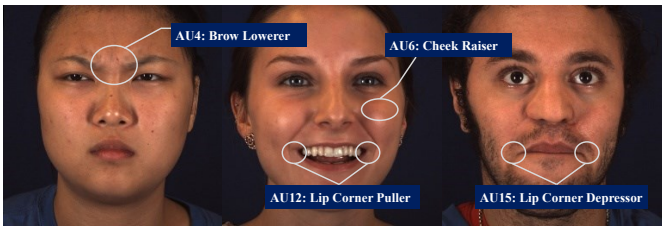


Fig. 1. Examples of AUs. AUs refer to the local movements of expressions, e.g. AU4 represents brow lowerer. AU6 represents cheek raiser. AU12 represents lip corner puller. AU15 represents lip corner depressor.

Since AU annotations require sophisticated expertise and are time-consuming, the size of annotated AU datasets is usually limited, especially in terms of identity variations, e.g. less than 50. As a result, most AU detection methods overfit to the training identities and do not generalize well to new subjects. To alleviate the overfitting problem on the small AU datasets, previous methods resort to various auxiliary information as regularization, including facial landmarks [33]–[36], unsupervised web images [37], emotion priors [38], textual AU descriptions [39] and so on. However, these additional constraints do not directly remove the interference of the training identities from the extracted visual features, thus limiting their performance.

Different from the previous works, we aim to make the first attempt to employ an expression embedding extracted from the in-the-wild expression dataset [40] without AU labels. The embedding can provide a strong prior for AU detection due to the two important properties: continuity and identity-independence. First, the embedding provides a continuous space for representing the fine-grained expressions. It is beneficial for AU detection since AUs usually show slight variations on the face. Second, the embedding is less sensitive to identities because the semantic similar expressions of different identities are analogous in the embedding space. This important property can be used to alleviate the overfitting problem in AU detection. Hence, our motivation is to leverage a continuous expression embedding space to represent AUs for accurate AU detection.

Driven by our motivation, we develop a novel AU detection framework aided by the Global-Local facial Expression Embedding, namely GLEE-Net. Our GLEE-Net consists of three branches that extract identity-independent facial expression

W. Zhang, L. Li, Y. Ding are with the Netease Fuxi AI Lab, Hangzhou, China.

W. Chen is with the Department of Hebei Agricultural University, Baoding, Hebei, China.

Z. Deng is with the Department of Computer Science, University of Houston, Houston, Texas, USA.

X. Yu is with the Department of the University of Queensland, Australia.

*Yu Ding is the corresponding author.

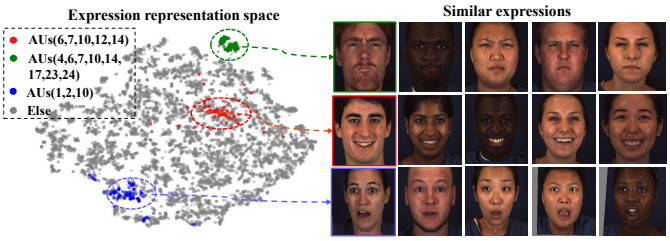


Fig. 2. T-SNE visualization of the distributions of some AU combinations in the embedding space of our GLEE-Net. Left: the same AU combinations distribute closely in the expression embedding space. Right: similar expressions in the expression embedding space often have similar AU labels and learning expressions can facilitate AU detection.

features for AU detection. To comprehend the overall facial expressions, we introduce a global branch that modeling the expression information as the deviation from the identity representation, which is inspired by DLN [59]. In this way, our global branch would be less sensitive to the identity information. However, such a structure lacks the perception of details of local regions. To resolve it, we design a local branch to focus on details of specific face regions.

In order to alleviate the problem of limited identities in AU datasets, different from existing methods with global and local branches, we first pretrain the two branches on an expression dataset [40] and then finetune them on the target AU dataset. In this manner, our network has seen various subjects' expressions although their AU labels are not available and moreover, acquires a compact expression embedding for AU detection. In Figure 2, we sample some images from BP4D and visualize their expression embeddings. As expected, the same AU combinations and similar expressions from different identities are close in our expression embedding space, thus facilitating AU classification.

Furthermore, in contrast to existing methods that only rely on auxiliary information of 2D images, we find that 3D facial information also provides important expression clues. Thus, we introduce a 3D global branch to obtain expression coefficients, as 3D expression features, through 3D face reconstruction. To fully exploit all the representations from our global-local branches, we design a Transformer-based multi-label classifier. Benefiting from the powerful global attention mechanism of Transformer [41], we can effectively fuse different representations and thus explore the correlations among multiple AUs. With the co-occurrence relationships of AUs, our network can predict AUs more accurately. Extensive experiments demonstrate that our approach significantly achieves superior performance on the widely-used DISFA, BP4D and BP4D+ datasets.

In summary, the contributions of our work are three-fold:

- We propose a novel Global-Local facial Expression Embedding Network (GLEE-Net) for AU detection, which can leverage additional facial expression data (without AU labels) to improve AU detection accuracy.
- We develop the global and local branches to extract the compact expression embeddings from face regions while paying attention to local facial details. To the best of our knowledge, our work is the first attempt to utilize continuous and compact expression features to represent AUs effectively. It achieves appealing generalization capability in addressing AU classification for unseen identities.
- We introduce a 3D global branch to extract expression coefficients through 3D face reconstruction for AU detection, and demonstrate that exploiting 3D face priors can further improve 2D AU detection.

II. Related Works

A. AU Detection with Auxiliary Information

The widely used AU datasets only contain limited subjects due to the difficulty of AU annotation, which is the main cause of overfitting. To resolve it, some works recourse to the various kinds of auxiliary information to enhance the model generalization and facilitate AU detection.

Introducing extra information of facial landmarks is a common practice in AU detection. To effectively extract the local features for AUs, JPML [33] utilized the landmarks to crop the facial patches instead of uniformly distributed grids. EAC-Net [34] also generated the spatial attention maps according to the facial landmarks and applied them to the different levels of networks. LP-Net [36] sent the detected facial landmarks into the P-Net to learn the person-specific shape information. JÂA-Net [35] proposed a multi-task framework combining landmark detection and AU detection. Besides this, there exist some other kinds of auxiliary information. Zhao et al. [37] utilized the unlabelled large-scale web images and proposed a weakly-supervised spectral embedding for AU detection. Cui et al. [38] constructed an expression-AUs knowledge prior based on the existing anatomic and psychological research and introduced the expression recognition model for AU detection. A series works [42]–[44] combined the tasks of AU detection and discrete expression recognition. SEV-Net [39] introduced the pre-trained word embedding to learn spatial attention maps based on the textual descriptions of AU occurrences. Xiang et al. [45] which introduced the temporal information, proposed an elegant linear model to untangle facial actions from a mixture of linearly-representable attributes. The aforementioned methods all directly or indirectly introduce additional data for producing extra regularization in AU detection. We propose to utilize the expression embedding as auxiliary information, which better improves the generalization capability of the AU detection.

B. AU Detection with Global and Local Features

Due to the local definition of AUs, many methods attempt to combine the full and regional facial features for AU detection. These works can be classified into three categories: patch-based, multi-task, and text-based methods.

Patch-based methods usually crop the full face into patches according to the local definitions of AUs. DSIN [46] cropped 5 patches from a full face based on landmarks and fed them with the full face into networks for learning the global and

local features. ROI-Net [47] also designed a prior landmark cropping rule to crop the inner feature maps. These methods usually suffer from performance degradation in the wild due to erroneous landmark estimation.

To facilitate the model with local features, multi-task methods combined AU detection with landmark detection [35], [48] or landmark-based attention map prediction [49]. In this way, models can extract global features from full faces and also focus on local details from the landmarks for better AU detection. However, these methods ignored that landmarks also contain rich identity information [50] which may aggravate the identity overfitting. SEV-Net [39] proposed to utilize the textual descriptions of local details to generate a regional attention map. In this way, it highlighted the local parts of the global features. However, it required the extra annotations for the descriptions. In addition, the global features of the previous works did not take the removal of the identity disturbance into account.

Different from the above works, our carefully-designed global branch is dedicated to eliminating identity disturbance, and our cropped patches for the local branch are based on positioned face patches instead of landmarks.

C. Expression Representations

The action units reflect the facial expression information and the model perception of the expression plays a crucial role in AU detection. The expression representation can be used to evaluate the expression perception capability of model. A common practice to represent expression is mapping the face images into a low-dimensional manifold, which describes the expressions without disturbance of identity, pose or illumination. Early works utilized the hidden features of the last or penultimate layer of the model trained in discrete expression classification tasks [4], [6], [51] as the expression representation, in which the extracted expression information reflects more information of the limited expression categories but neglect the complicated and fine-grained facial expressions.

Different from them, a compact and continuous embedding for representing facial expressions was proposed by Vemulapalli and Agarwala [40]. It constructed a large-scale facial dataset annotated with expression similarity in a triplet way. Through a large number of triplet comparisons, the trained expression embedding can perceive slight expression changes.

To further reduce the identity influence, Zhang et al. [59] developed a Deviation Learning Network (DLN) with a two-branch structure to achieve more compact and smooth expression embedding. Taha et al. [52] proposed a generic multimodal mesh surface representation that incorporates both 2D and 3D information. The extensive experiments proved that the representation can facilitate AU detection and expression recognition. 3D Morphable Model (3DMM) [53], [54] has been proposed to fit identities and expression parameters from a single face image. Expressions are represented as the coefficients of predefined blendshapes in 3DMM. The estimated expression coefficients are then used for talking head synthesis [9], [55], expression transfer [56], [57] or face manipulation [58].

III. Method

The architecture of the proposed GLEE-Net is shown in Figure 3, which takes an image as input and outputs a binary vector to indicate the occurrence of each AU. The whole framework consists of a global branch, a local branch, a 3D global branch and a Transformer classifier. The global branch extracts the full face feature to model the full face expression while the local branch focuses on detailed local information. The two branches are pretrained on the FEC expression dataset [40] and then finetuned on the AU dataset to alleviate the issue of limited identities. To further enrich 2D facial representations, the 3D global branch extracts the expression coefficients through 3D face reconstruction. Finally, the Transformer classifier carries out the final AU detection from the combined features of three branches with the powerful attention mechanism.

A. Global Branch

Inspired by DLN [59], the global branch models the expression feature vector V_{exp} as the deviation from the identity vector V_{id}. Specifically, the global branch consists of two siamese models, i.e., the face model and the identity model. The identity model and the face model are initialized with the pretrained FaceNet [60] for a face recognition task [61]. Then, we freeze the identity model and train the face model to learn the expression deviation. The extracted full face expression feature vector V_{exp} is obtained by:

$$V_{exp} = V_{face} - V_{id}. \tag{1}$$

At the beginning of training the expression embedding, the Identity model and the Face model share the same initialization model parameters, but the setting of batchnorm parameters in the Face model is randomly initialized. The distinguishable batchnorm parameters between the Identity model and the Face model ensure that V_{exp} is not a zero vector at the beginning of the training, which guarantees valid gradient propagation in training. On the other hand, as the training progresses, the parameters of the Face model will increasingly diverge from the Identity model to represent identity-invariant facial expressions with the deviation vectors V_{exp}. In other words, V_{exp} will be valued to represent facial expression by comparing the outputs of the frozen Identity model and trainable Face model. The Face model is updated to satisfy the labels of triplet facial expressions, which makes V_{exp} away from zero.

The deviation model of the global branch benefits from an effective feature initialization that can alleviate the disturbance of expression-irrelevant information, such as identity, pose, etc. After a linear layer for dimension reduction, we obtain G_{exp} as the global expression feature vector.

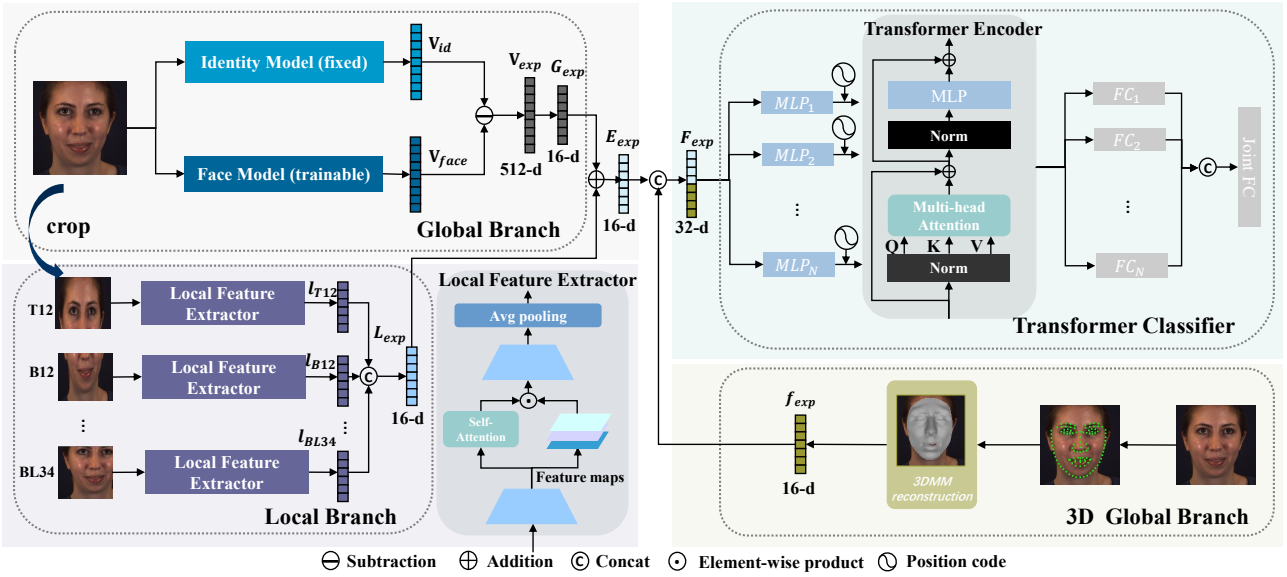


Fig. 3. Pipeline of our proposed AU detection framework GLEE-Net. The global and local branches extract the full and partial facial information and are pretrained on the FEC dataset to provide an effective expression prior knowledge for AU detection. To further enhance the expression representation, we exploit the expression coefficients from a 3D global branch. The Transformer classifier models the diverse correlations between AUs and then predicts AUs.

B. Local Branch

Different from DLN, we introduce a local branch to facilitate the global branch with more detailed local information, which is also beneficial for AU detection due to the local nature of AUs. First, we crop the image into 16 parts for local part extraction. Since the expression dataset contains a large number of in-the-wild images, it is hard to locate specific face regions accurately. Therefore, we choose to crop the image according to the whole image area instead of facial landmarks. Specifically, we crop three-quarters of the image from left, right, top and bottom and call them L34, R34, T34 and B34 respectively. Similarly, we crop half from the left, right, top and bottom as L12, R12, T12 and B12 respectively. We also crop the image from two directions simultaneously to construct eight local parts as TL34, TR34, TL12, TR12, BL34, BR34, BL12 and BR12 respectively. Figure 4 shows some examples of our cropping strategy. We can observe that different parts focus on different face regions (e.g., T12 focuses on eyebrows and eyes).

After obtaining the local parts of a face, we resize them to 96×96 and feed them into a group of local feature extractors separately. The structure of the local feature extractor is illustrated in Figure 3. To assign the importance weight for different local parts, there exist some common strategies, such as VLAD [62], VLAAD [63] and self-attention [64]. Among these strategies, we choose to use self-attention because it is easy to implement and has excellent performance. Specifically, after four 3×3 convolutional layers, a self-attention module is employed to strengthen the importance of crucial local positions on a face.

Afterwards, the feature maps are resized through average pooling. We concatenate the local features of the 16 face regions and resize the final local expression feature L_{exp} to 16 by MLP layers. We obtain the final expression embedding E_{exp} by adding L_{exp} and the global expression feature G_{exp}. E_{exp} has the comprehensive expression representation capability from the global-local 2D face regions.

Fig. 4. Illustration of cropped facial parts. Four of the 16 specific parts used in our method are shown (L34, BL34, T12 and BR12).

C. 3D Global Branch

To further enrich the expression embedding, the 3D global branch estimates the expression coefficients from a 3DMM. The 3D expression coefficients also reflect the AU movement information and our work is the first to introduce it as another input feature for AU detection.

Given the face shape coefficients $f_s \in \mathbb{R}^{N_s}$ and the expression coefficients $f_{exp} \in \mathbb{R}^{N_e}$, the parametric 3D mesh $M(s, e)$ is represented as:

$$M(s,e) = M_0 + \sum_{i=1}^{N_s} S_i \cdot (f_s)_i + \sum_{j=1}^{N_e} E_j \cdot (f_{exp})_j, \quad (2)$$

where M_0 is the mean face mesh, and $\{S_i\}_{i=1}^{N_s}$ and $\{E_i\}_{i=1}^{N_e}$ are the linear shape and expression bases, respectively. We take M_0 and $\{S_i\}_{i=1}^{N_s}$ with $N_s = 60$ from LSFM [54] and

normalize them into the range of $[-3, 3]$. We sculpt $N_e = 51$ blendshapes on LSFM as $\{E_i\}_{i=1}^{N_e}$.

To reconstruct f_{exp} from the input image, we perform 3DMM fitting by minimizing the energy function:

$$\mathcal{L}_{3D} = \sum_{i=1}^{N_l} D(p_i, P(M(f_s, f_{exp})_i, \xi) + \lambda_e \left\| f_{exp} \right\|_2^2 + \lambda_s \left\| f_s \right\|_2^2, \tag{3}$$

where $N_l = 68$ is the number of facial landmarks, p_i is the pixel coordinate of the detected i-th landmark, and $M(s, e)_i$ is the 3D vertex of the i-th landmark on mesh M. P is the perspective projection matrix [65] that projects a 3D vertex into its homogeneous pixel coordinate according to the head pose ξ. The first term $D(\cdot, \cdot)$ measures the distance between a detected image landmark and its corresponding projected mesh landmark. Since the contour landmark (e.g., lip contours) distributions of the image detector and 3D mesh notation are not perfectly aligned, for non-endpoint landmarks, we define the point-to-curve pixel distance between the projected 3D landmark point and the detected 2D face contour curve. The second and third regularization terms with weights $\lambda_e = 10^{-4}$ and $\lambda_s = 10^{-4}$ respectively. We employ the Levenberg-Marquard (LM) algorithm [66] to solve the best f_{exp} and f_s.

After optimization, the estimated expression coefficients f_{exp} are concatenated with the global-local expression embedding E_{exp} for AU classification. f_{exp} is not applied in the pretraining process of expression embedding because the quality of in-the-wild FEC [40] images is often worse than that of AU detection images and may lead to estimation errors in 3D reconstruction.

D. Transformer Classifier

We notice that the correlation among multiple AUs provides additional constraints and thus facilitates AU predictions. Therefore, rather than a conventional multi-class classifier, we model the correlation among AUs by Transformer encoders [41]. In this way, the prediction of the occurrence of each AU takes into account the correlations of all the other AUs. As shown in Figure 3, the parallel MLP layers first extract multi-view information related to AUs from the joint expression representation F_{exp}. Then, three Transformer encoders take the extracted multi-view AU features as an input sequence. Based on the multi-head attention mechanism, the Transformer encoders can fully exploit the latent correlations among AUs. Finally, the enhanced features are passed through independent FC layers as initial predictions and a joint FC layer is employed to fuse the initial predictions to achieve a more accurate final prediction.

E. Network Training

Pretraining for Expression Embedding: The global and local branches are pretrained on the expression dataset FEC [40] with a triplet loss to obtain a compact expression embedding E_{exp} as an auxiliary task. This is different from DLN since it only focuses on the global face. FEC is a large-scale in-the-wild expression dataset that contains 500,203 triplets annotated by human perception of expression.

We also utilize the hard sample mining strategy from [67] which introduces a hierarchical labeling approach to find the more valuable triplets. Because of the pyramid structure of the hierarchical annotation design, we can automatically generate more triplet results by chain comparisons.

One triplet consists of one Anchor (A), Positive (P) and Negative (N). A and P have more similar expressions than N. Following such annotation, we use the triplet loss to pretrain the two branches as follows:

$$\begin{aligned}\mathcal{L}_{tri} = &\max(0, \|E_{exp}^A - E_{exp}^P\|_2^2 - \|E_{exp}^A - E_{exp}^N\|_2^2 + m) \\ &+ \max(0, \|E_{exp}^A - E_{exp}^P\|_2^2 - \|E_{exp}^P - E_{exp}^N\|_2^2 + m),\end{aligned} \tag{4}$$

where $E_{exp}^{(\cdot)}$ represents the corresponding expression embedding after normalization, and m is the margin. Due to a large number of expression comparisons from different identities in FEC, the extracted expression embeddings can be identity-independent and continuous. Also, the expression representation can be more robust with the local branch since the cropped parts provide rich detailed local information. The expression embedding is of great value for the generalization of AU detection.

Training for AU Detection: Aided by the effective expression embeddings, we finetune the whole network for the AU detection task. Specifically, we employ the weighted cross entropy loss to train our AU classifiers, defined by:

$$\mathcal{L}_{AU} = -\frac{1}{N_a} \sum_{i=1}^{N_a} \frac{1}{r^i} [g^i \log p^i + (1 - g^i) \log(1 - p^i)], \tag{5}$$

where N_a is the number of AUs, g^i is the ground-truth binary label of the i-th AU, p^i is the predicted value, and r^i is the prior occurrence ratio of the i-th AU in the training set to balance the weight of different AUs. The loss is applied on both the initial predictions before joint FC and the final predictions. In the inference stage, the i-th AU is determined as presence if $p^i > 0.5$.

IV. Experiment

A. Datasets

We evaluate our method on three widely-used AU detection datasets: BP4D [77], DISFA [78], and BP4D+ [79]. We divide the training and validation set based on the subject-exclusive rule, which means the identities in validation set cannot appear in the training set. The following experiments report all the available AUs in these datasets. Currently, there is no dataset that provides annotations for all AUs defined in the FACS. Besides, the dataset division in all the following experiments is based on identity information to ensure that the training set and testing set do not have images of the same identity.

BP4D is composed of about 140,000 frames from 328 video clips. Each video shows the spontaneous reactions of a subject when he is doing a specific task. The subjects contain 23 females and 18 males. The annotation of each frame includes

TABLE I
COMPARISONS OF OUR METHOD AND THE STATE-OF-THE-ART METHODS ON BP4D IN TERMS OF F1 SCORES (%). THE BEST RESULTS ARE SHOWN IN BOLD, AND THE SECOND BEST ARE IN BRACKETS.

Method	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU14	AU15	AU17	AU23	AU24	Avg.
LSVM [68]	23.2	22.8	23.1	27.2	47.1	77.2	63.7	64.3	18.4	33.0	19.4	20.7	35.3
JPML [33]	32.6	25.6	37.4	42.3	50.5	72.2	74.1	65.7	38.1	40.0	30.4	42.3	45.9
DRML [69]	36.4	41.8	43.0	55.0	67.0	66.3	65.8	54.1	33.2	48.0	31.7	30.0	48.3
EAC-Net [34]	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
DSIN [46]	51.7	40.4	56.0	76.1	73.5	79.9	85.4	62.7	37.3	62.9	38.8	41.6	58.9
ARL [70]	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 [71]	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
LP-Net [36]	43.4	38.0	54.2	77.1	76.7	83.8	87.2	63.3	45.3	60.5	48.1	54.2	61.0
UGN-B [72]	54.2	46.4	56.8	76.2	76.7	82.4	86.1	64.7	51.2	63.1	48.5	53.6	63.3
JÂA-Net [35]	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
MHSA-FFN [49]	51.7	49.3	**61.0**	77.8	79.5	82.9	86.3	[67.6]	51.9	63.0	43.7	[56.3]	64.2
SEV-Net [39]	[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
HMP-PS [73]	53.1	46.1	56.0	76.5	76.9	82.1	86.4	64.8	51.5	63.0	**49.9**	54.5	63.4
MAL [42]	47.9	[49.5]	52.1	77.6	77.8	82.8	88.3	66.4	49.7	59.7	45.2	48.5	62.2
ME-GraphAU [74]	52.7	44.3	60.9	77.9	[80.1]	85.3	**89.2**	**69.4**	[55.4]	[64.4]	[49.8]	55.1	[65.5]
KDSSRL [75]	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
GLEE-Net (Ours)	**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**

TABLE II
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.
LSVM [68]	10.8	10.0	21.8	15.7	11.5	70.4	12.0	22.1	21.8
DRML [69]	17.3	17.7	37.4	29.0	10.7	37.7	38.5	20.1	26.7
APL [76]	11.4	12.0	30.1	12.4	10.1	65.9	21.4	26.9	23.8
EAC-Net [34]	41.5	26.4	66.4	50.7	8.5	**89.3**	88.9	15.6	48.5
DSIN [46]	42.4	39.0	68.4	28.6	46.8	70.8	90.4	42.2	53.6
ARL [70]	43.9	42.1	63.6	41.8	40.0	76.2	[95.2]	66.8	58.7
SRERL [71]	45.7	47.8	56.9	47.1	45.6	73.5	84.3	43.6	55.9
LP-Net [36]	29.9	24.7	72.7	46.8	49.6	72.9	93.8	56.0	56.9
UGN-B [72]	43.3	48.1	63.4	49.5	48.2	72.9	90.8	59.0	60.0
JÂA-Net [35]	**62.4**	**60.7**	67.1	41.1	45.1	73.5	90.9	[67.4]	63.5
MHSA-FFN [49]	46.1	48.6	72.8	**56.7**	50.0	72.1	90.8	55.4	61.5
SEV-Net [39]	55.3	53.1	61.5	53.6	38.2	71.6	**95.7**	41.5	58.8
HMP-PS [73]	38.0	45.9	65.2	50.9	50.8	76.0	93.3	**67.6**	61.0
MAL [42]	43.8	39.3	68.9	47.4	48.6	72.7	90.6	52.6	58.0
ME-GraphAU [74]	54.6	47.1	[72.9]	[54.0]	**55.7**	76.7	91.1	53.0	63.1
KDSSRL [75]	60.4	[59.2]	67.5	52.7	51.5	76.1	91.3	57.7	[64.5]
GLEE-Net (Ours)	[61.9]	54.0	**75.8**	45.9	**55.7**	[77.6]	92.9	60.0	**65.5**

the occurrence or absence of 12 AUs and the intensity of 5 AUs. For fair comparisons, we choose the same three-fold dataset division as the previous works [34], [35] and perform the three-fold cross validation.

DISFA is composed of about 130,000 frames from 27 video clips. Like BP4D, the expressions in videos come from human spontaneous reactions. The subjects contain 12 females and 15 males, which is less than BP4D. The annotation of each frame includes the intensities (range from 0 to 5) of 8 AUs. Following previous works, the intensity with $\{0, 1\}$ value represents the AU absence and the intensity with $\{2, 3, 4, 5\}$ value represents the AU occurrence. We also compare our method with other works based on the three-fold cross validation.

BP4D+ extends more subjects based on BP4D. It contains 1400 video clips from 140 subjects (82 females and 58 males). About 198,000 frames are annotated with the occurrence or absence of the same 12 AUs in BP4D. Following the previous works, we also perform three-fold cross validation. Moreover, to further prove the superior generalization of our method, we also construct the experiment that is trained on BP4D and evaluated on BP4D+ following the works [35], [70].

B. Implementation Details

We first pretrain the global and local branches on FEC [40] and then finetune the whole framework on evaluated AU datasets. In the pretraining stage, we use an SGD optimizer with momentum 0.9 and learning rate 2×10^{-4}. The pretraining converges within 10 epochs with a batch size of 30. We set the margin m to 0.1 for one-class triplets and 0.2 for two-class and three-class triplets, as recommended by DLN [59]. One-class triplets involve three images from the same emotion tag, while two-class and three-class triplets contain three images from two or three different emotion tags. These triplet types are included in FEC annotations. We use Euclidean distance in the pre-training processing, following the DLN [59].

In the finetuning stage, we train the entire framework with an SGD optimizer with momentum 0.9 and a learning rate 2×10^{-3} for 10 epochs. We employ the method [80] as the facial landmark detector. The input face images are aligned by similarity transformation as JÂA-Net [35]. The computational cost of our method involves 8.72 GFLOPs. The number of parameters is 56.6M. Our GLEE-Net is not heavy-weighted because our local feature extractor for per patch only contains a few layers, and the input of the Transformer encoder is 12 AU features of dimension 32. For the expression embedding pre-training, each iteration takes an average of 0.2482 seconds, and each epoch takes approximately 45.18 minutes to complete. During the training process for AU detection, taking the BP4D dataset as an example, each step iteration takes an average of 0.1375 seconds, and each epoch takes approximately 5.502 minutes to complete. The time spent on other datasets is roughly the same. For the test stage, the inference time is relatively small, with only 0.0241 seconds needed to process a single batch. All time analyses are conducted using a single GPU of 3090Ti.

C. Comparison with the State-of-the-Art

We first compare our method with the state-of-the-art. The compared methods include LSVM [68], JPML [33], DRML [69], APL [76], EAC-Net [34], DSIN [46], ARL [70], SRERL [71], ML-GCN [82], MS-CAM [83], LP-Net [36], FACS3D-Net [81], UGN-B [72], JÂA-Net [35], MHSA-FFN [49], SEV-Net [39], HMP-PS [73], MAL [42], ME-GraphAU [74], KDSSRL [75]. Following the experimental settings of most previous works, we employ the F1 score [84] on the frame level as the evaluation metric.

Evaluation on BP4D. The results on BP4D are shown in Table I. The average F1 score of our method is 65.7%, which outperforms all the competing methods. In terms of the performance on single AUs, our GLEE-Net achieves the highest or the second highest F1 score among all the methods on 6 of the 12 evaluated AUs. Compared to SEV-Net, JÂA-Net and MAL that employ extra auxiliary information to facilitate AU detection, our method obtains better results with the compact expression embedding as prior knowledge. Our method also outperforms MHSA-FFN which also introduces Transformer into classifiers. ME-GraphAU achieves a close result as our method on BP4D but shows a significant disadvantage on DISFA.

Evaluation on DISFA. Table II shows the results on DISFA. The average F1 score of our method outperforms all the state-of-the-art methods. Compared to the most recent works SEV-Net, HMP-PS, MHSA-FFN and ME-GraphAU, our method obtains the improvement of 6.7%, 4.5%, 4.0%, and 2.4%, respectively. As the number of subjects in DISFA is smaller than that of BP4D, most of the previous methods suffer from overfitting to the appearance of training subjects more severely. Therefore, benefiting from our identity-independent expression representations, our method achieves an even more significant improvement.

Evaluation on BP4D+. Table III shows the results of the three-fold cross-validation on BP4D+. Results of the compared methods are reported by SEV-Net [39]. Our method achieves the F1 score of 63.7% and outperforms all the state-of-the-art methods. To further evaluate the generalization performance on large-scale testing identities, we perform the cross-dataset validation. We do not choose the BP4D and DISFA since they have different labeled AU categories. Therefore, we train our method on BP4D and evaluate it on the full BP4D+ of 140 subjects. The results are shown in Table IV. We use the reported results in the work [35] for comparison. Again, our method outperforms all the compared methods under the large-scale cross-dataset evaluation. It proves that our method can extract identity-independent information and generalize well to new identities.

Through comparisons on different datasets, we observe that our method significantly improves the F1 scores for some AUs, but not for others. This is because our work makes efforts to alleviate data sparsity and imbalance and improve generalization. Our approach exhibits clear superiority, better generalization, and less susceptibility to imbalanced data distribution when observing those AUs with insufficient data samples, including AU1, AU4, AU17, and AU24 in BP4D, AU4 and AU9 in DISFA, and AU1, AU2, and AU24 in BP4D+. On the other hand, the other AUs have relatively-sufficient data, our method shows comparable performance or relatively weak improvement.

D. Ablation Study

We also conduct an ablation study to evaluate the effectiveness of each module by removing or replacing it with a baseline method. The results are shown in Table V.

Prior Knowledge. In the experiment without pretraining on the expression dataset (w/o pretrain), we can see an apparent drop in average F1. It proves the significance of prior knowledge from the compact expression embedding. We also set up an experiment with fixed pretrained parameters of global and local branches (w. fixed GB & LB) and obtain a lower average F1, which indicates a domain gap exists between FEC and BP4D datasets.

Expression Representation. To validate the effectiveness of representations extracted by the 3D global, global and local branches, we construct the model with only 3D global branch (w/o GB & LB) and only the global and local branches (w/o 3DGB) individually. To indicate the significance of the global and local branches, we replace these two branches with a ResNet [85] (w. ResNet). Also, we build up a model without the local branch (w/o LB) to confirm the usefulness of the local branch. The experiments show removing those components lead to degraded performance compared to full GLEE-Net, proving that three branches are all effective. The local branch is used to supply extra local detail information to the global branch rather than working individually.

Classifier. We replace the Transformer classifier with an MLP classifier (w/o TC), the decrease of average F1 indicates that the Transformer classifier improves AU detection

TABLE III
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 [81]	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 [82]	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 [83]	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 [39]	[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]
GLEE-Net (Ours)	**54.2**	**46.3**	[38.1]	[86.2]	87.6	[90.4]	**89.5**	81.3	46.3	[47.4]	57.6	**39.6**	**63.7**

TABLE IV
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 [34]	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 [70]	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 [35]	[39.7]	35.6	30.7	[82.4]	[84.7]	[88.8]	[87.0]	[62.2]	38.9	**46.4**	**48.9**	**36.0**	[56.8]
GLEE-Net (Ours)	**39.8**	**37.9**	**41.6**	**83.4**	**88.2**	**90.2**	**87.4**	**76.6**	**48.3**	42.9	47.7	29.8	**59.5**

TABLE V
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	54.8	44.5	52.3	77.6	75.9	82.5	86.7	63.6	45.6	62.3	47.5	51.0	62.0
w. fixed GB & LB	55.1	44.9	49.1	76.5	75.9	82.2	87.2	65.0	36.3	59.8	34.4	50.2	59.7
w. ResNet	55.0	45.2	52.0	77.8	77.8	83.3	[87.8]	**65.8**	46.2	63.8	45.8	53.5	62.8
w/o GB & LB	55.2	45.2	46.1	76.1	75.6	81.5	86.0	64.8	36.0	59.7	34.1	49.4	59.1
w/o 3DGB	54.4	43.3	59.2	[79.5]	[78.1]	[84.7]	87.6	64.6	[53.9]	63.9	47.3	[58.2]	64.6
w/o LB	58.9	[46.8]	55.9	78.3	78.0	83.3	87.6	62.1	52.8	[64.2]	**50.3**	53.5	64.3
w/o TC	[59.7]	43.0	**64.0**	78.7	77.2	83.3	87.5	63.8	**54.3**	[64.2]	46.7	56.0	[64.9]
w/o joint FC	56.3	**47.3**	57.0	77.9	76.1	84.0	87.7	[65.4]	51.6	63.2	[48.9]	56.5	64.3
GLEE-Net (Full)	**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**

performance. We also remove the joint FC layer of the Transformer classifier and the final prediction (w/o joint FC). The performance drop validates the practical design of our two-stage predictions. In comparison to other modules, the Transformer Classifier (w/o TC) and joint FC have a relatively minor impact. This could be because our feature extraction is well-designed and places low demands on the tail-end classification network.

E. Visual Analysis

In this part, we present some t-SNE visualization results to investigate the impact of our proposed expression embedding pre-training and each feature branches used in our pipeline.

Figure 5 presents the t-SNE results of E_{exp} obtained through the use of our pre-trained expression embedding (Ours) and the absence of pre-training (w/o pre-train). We randomly select five AU labels from BP4D, and plot their respective t-SNE distributions. The legends in Fig. 5, such as "AUs(6,7,10,12)," denote the occurrence of AUs 6, 7, 10, and 12 in the respective samples, and follow the same convention as the other legends. From the figure, it is evident that our expression embedding pretraining method allows the model to learn a more reasonable feature distribution. Specifically, samples with the same label are clustered closer together, and samples with similar AUs (e.g., AUs(6,7,10,12,14) and AUs(6,7,10,12)) are closer to each other than to samples with different labels.

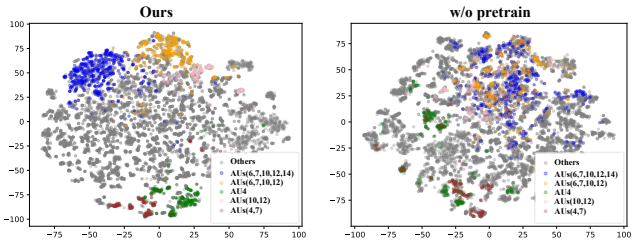


Fig. 5. T-sne result of E_{exp} obtained by using the pretraining (Ours) and not using pretraining (w/o pretrain). We randomly choose five AU combinations in BP4D and it can be seen that our expression embedding pretraining lead to a more reasonable distribution in the feature space.

To demonstrate the effectiveness of each feature branch of GLEE-Net, we sample some images with seven common AU combinations in BP4D, and show the t-SNE visualization of their expression representations F_{exp} in Figure 6. We compute the expression representation variance for each AU combination and average them as *Ave-Var*. Samples of the same AU combination from different identities are distributed more closely under our full method (GB & LB & 3DGB) than other ablation settings. This indicates that our expression representation focuses on the AU features rather than over-

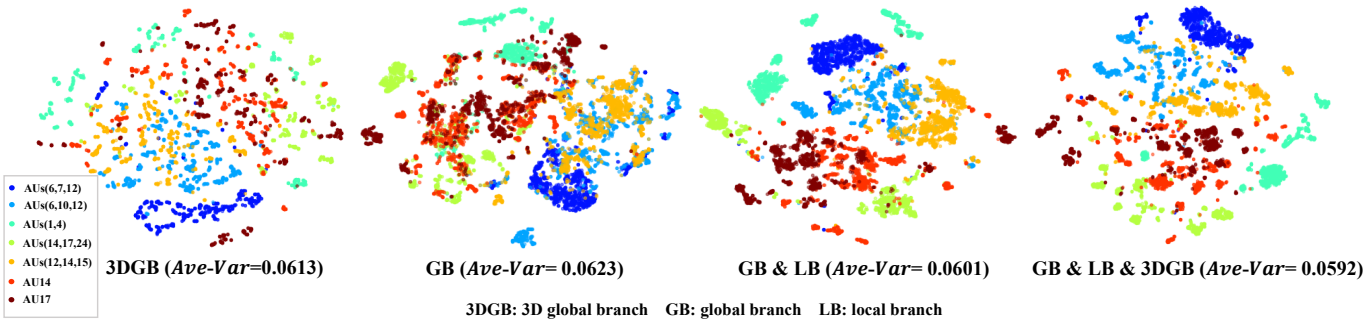


Fig. 6. T-SNE visualization of sampled AU combinations in the four different representation spaces. It is observed that representations from our proposed GLEE-Net (GB & LB & 3DGB) are more compact in each AU combination. *Ave-Var* is the average of representation variances of all the AU combinations.

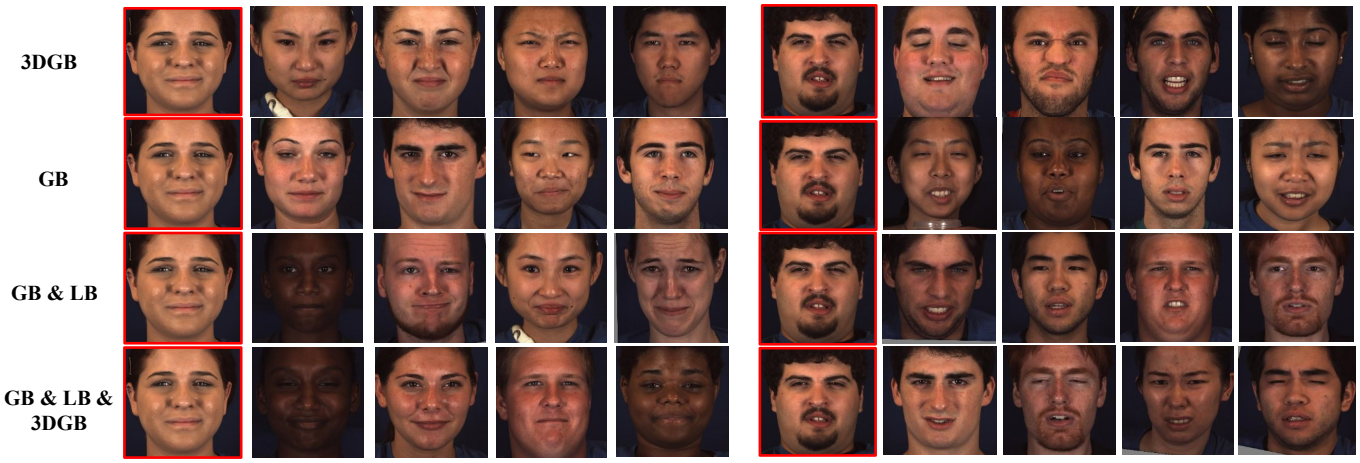


Fig. 7. Illustration of our expression representation facilitating AU detection. We attain one image as a center (marked in red) and sample its surrounding features from the expression representation space. It can be seen that images with similar AUs are retrieved in this space.

fitting to training identities. Figure 7 shows the expression representation facilitates AU detection. We choose one image as center (in red) and sample its surrounding features in the representation space. Under the joint features of the three branches (GB & LB & 3DGB), neighbouring images all have the same facial expressions. The other three versions roughly capture similar facial actions, but there are still subtle variances of AUs.

F. Discussion

Existing top-ranked works like MHSA-FFN [49] and SEV-Net [39] also use additional dataset information, such as landmarks, word embedding, and so on, but they still suffer from the identity overfitting problem on DISFA that contains smaller subjects. Compared with previous works, our method is the first to introduce the effective expression embedding and 3D expression coefficients to alleviate the issue of overfitting caused by limited identities of AU datasets. Our method makes full of the identity-independent expression representation based on additional dataset information (without AU labels) to facilitate AU detection.

Compared with MHSA-FFN and SEV-Net, our method achieves the promising results on widely-used BP4D, BP4D+ and DISFA datasets. Also, cross-dataset evaluation demonstrates our better generalization than the state-of-the-art methods.

In terms of the limitation of our method, we pre-extract the 3DMM features to improve training and prediction efficiency. However, this approach may not be suitable for real-world scenarios as it can affect the speed of real-time detection. Despite the aforementioned limitation, our method exhibits outstanding performance in terms of results. In our future works, we intend to seek methods to overcome these limitations and investigate in-the-wild AU datasets such as SEWA [86] and Aff-wild2 [87], [91], which have not been widely employed in existing works. We will exploit the effect of different orders of AU learning on Transformer-based models' performance. And we will investigate the potential improvement by using the semi-supervised method [88] and VLAAD [63]. Also, the methods used in [30], [31], [89], [90] are also worth paying attention to.

V. Conclusion

This paper presents a novel AU detection framework, dubbed GLEE-Net, by fully taking advantage of available facial expression data. Through the global and local branches,

our GLEE-Net extracts rich representations of input faces in a compact and continuous expression space. Taking advantage of the expression representations, similar AU combinations cluster closer in the latent space. This not only significantly facilitates AU classification but also makes AU detection generalized well to various subjects. Moreover, our introduced 3D facial expression branch provides complementary information to 2D AU detection. Our Transformer-based classifier effectively fuses representations of three branches and produces more accurate AU detection results than the state-of-the-art.

VI. Acknowledgement

This work is supported by the 2022 Hangzhou Key Science and Technology Innovation Program (No. 2022AIZD0054), the Key Research and Development Program of Zhejiang Province (No. 2022C01011) and the National Key R&D Program of China (No. 2022YFF09022303). This research is funded in part by ARC-Discovery grant (DP220100800 to XY) and ARC-DECRA grant (DE230100477 to XY). We thank all anonymous reviewers and ACs for their constructive suggestions.

References

[1] R. Vemulapalli and A. Agarwala, "A compact embedding for facial expression similarity," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5683–5692.

[2] S. Li and W. Deng, "Blended emotion in-the-wild: Multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning," *International Journal of Computer Vision*, vol. 127, no. 8, pp. 884–906, 2019.

[3] K. Chen, X. Yang, C. Fan, W. Zhang, and Y. Ding, "Semantic-rich facial emotional expression recognition," *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 1906–1916, 2022.

[4] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.

[5] Z. Wen, W. Lin, T. Wang, and G. Xu, "Distract your attention: Multi-head cross attention network for facial expression recognition," *arXiv preprint arXiv:2109.07270*, 2021.

[6] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 4057–4069, 2020.

[7] A. P. Fard and M. H. Mahoor, "Ad-corre: Adaptive correlation-based loss for facial expression recognition in the wild," *IEEE Access*, vol. 10, pp. 26 756–26 768, 2022.

[8] Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, and D. Li, "Makelttalk: speaker-aware talking-head animation," *ACM Transactions On Graphics (TOG)*, vol. 39, no. 6, pp. 1–15, 2020.

[9] Z. Zhang, L. Li, Y. Ding, and C. Fan, "Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3661–3670.

[10] H. Zhou, Y. Sun, W. Wu, C. C. Loy, X. Wang, and Z. Liu, "Pose-controllable talking face generation by implicitly modularized audio-visual representation," in *IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 4176–4186.

[11] S. Wang, L. Li, Y. Ding, C. Fan, and X. Yu, "Audio2head: Audio-driven one-shot talking-head generation with natural head motion," in *International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, 8 2021, pp. 1098–1105.

[12] S. Wang, L. Li, Y. Ding, and X. Yu, "One-shot talking face generation from single-speaker audio-visual correlation learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 2531–2539.

[13] X. Ji, H. Zhou, K. Wang, Q. Wu, W. Wu, F. Xu, and X. Cao, "Eamm: One-shot emotional talking face via audio-based emotion-aware motion model," in *ACM SIGGRAPH 2022 Conference Proceedings*, 2022, pp. 1–10.

[14] Z. Zhang and Y. Ding, "Adaptive affine transformation: A simple and effective operation for spatial misaligned image generation," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 1167–1176.

[15] B. Liang, Y. Pan, Z. Guo, H. Zhou, Z. Hong, X. Han, J. Han, J. Liu, E. Ding, and J. Wang, "Expressive talking head generation with granular audio-visual control," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3387–3396.

[16] Y. Ma, S. Wang, Z. Hu, C. Fan, T. Lv, Y. Ding, Z. Deng, and X. Yu, "Styletalk: One-shot talking head generation with controllable speaking styles," *arXiv preprint arXiv:2301.01081*, 2023.

[17] L. Chen, G. Cui, C. Liu, Z. Li, Z. Kou, Y. Xu, and C. Xu, "Talking-head generation with rhythmic head motion," in *ECCV*, 2020, pp. 35–51.

[18] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, "Learning affective features with a hybrid deep model for audio–visual emotion recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 3030–3043, 2018.

[19] R. Chen, W. Zhou, Y. Li, and H. Zhou, "Video-based cross-modal auxiliary network for multimodal sentiment analysis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 12, pp. 8703–8716, 2022.

[20] F. Qiu, W. Kong, and Y. Ding, "Intermulti: Multi-view multimodal interactions with text-dominated hierarchical high-order fusion for emotion analysis," *arXiv preprint arXiv:2212.10030*, 2022.

[21] Y. Ou, Z. Chen, and F. Wu, "Multimodal local-global attention network for affective video content analysis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 5, pp. 1901–1914, 2021.

[22] J. Tang, D. Liu, X. Jin, Y. Peng, Q. Zhao, Y. Ding, and W. Kong, "Bafn: Bi-direction attention based fusion network for multimodal sentiment analysis," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2022.

[23] K. Zhang, Y. Li, J. Wang, E. Cambria, and X. Li, "Real-time video emotion recognition based on reinforcement learning and domain knowledge," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1034–1047, 2022.

[24] J. Guo, J. Tang, W. Dai, Y. Ding, and W. Kong, "Dynamically adjust word representations using unaligned multimodal information," in *ACM International Conference on Multimedia*, 2022, pp. 3394–3402.

[25] F. Qiu, C. Xie, Y. Ding, and W. Kong, "Effmulti: Efficiently modeling complex multimodal interactions for emotion analysis," *arXiv preprint arXiv:2212.08661*, 2022.

[26] P. Ekman and W. Friesen, "Facial action coding system: A technique for the measurement of facial movement," *Consulting Psychologists Press Palo Alto*, vol. 12, 01 1978.

[27] M. Pantic and L. Rothkrantz, "Facial action recognition for facial expression analysis from static face images," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 34, pp. 1449 – 1461, 07 2004.

[28] G. Zhao and X. Li, "Automatic micro-expression analysis: Open challenges," *Frontiers in Psychology*, vol. 10, 08 2019.

[29] Y. Zhang, Q. Ji, Z. Zhu, and B. Yi, "Dynamic facial expression analysis and synthesis with mpeg-4 facial animation parameters," *IEEE Transactions on circuits and systems for video technology*, vol. 18, no. 10, pp. 1383–1396, 2008.

[30] J. Tang, Z. Li, H. Lai, L. Zhang, S. Yan *et al.*, "Personalized age progression with bi-level aging dictionary learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 905–917, 2017.

[31] X. Shu, J. Tang, H. Lai, L. Liu, and S. Yan, "Personalized age progression with aging dictionary," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3970–3978.

[32] D. R. Rubinow and R. M. Post, "Impaired recognition of affect in facial expression in depressed patients," *Biological psychiatry*, vol. 31, no. 9, pp. 947–953, 1992.

[33] K. Zhao, W.-S. Chu, F. De la Torre, J. F. Cohn, and H. Zhang, "Joint patch and multi-label learning for facial action unit detection," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[34] 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.

[35] Z. Shao, Z. Liu, J. Cai, and L. Ma, "Jaa-net: joint facial action unit detection and face alignment via adaptive attention," *International Journal of Computer Vision*, vol. 129, no. 2, pp. 321–340, 2021.

[36] X. Niu, H. Han, S. Yang, Y. Huang, and S. Shan, "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*, 2019, pp. 11 917–11 926.

[37] K. Zhao, W.-S. Chu, and A. M. Martinez, "Learning facial action units from web images with scalable weakly supervised clustering," in *Proceedings of the IEEE Conference on computer vision and pattern recognition*, 2018, pp. 2090–2099.

[38] 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, 2020.

[39] 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.

[40] R. Vemulapalli and A. Agarwala, "A compact embedding for facial expression similarity," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5683–5692.

[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[42] Y. Li and S. Shan, "Meta auxiliary learning for facial action unit detection," *IEEE Transactions on Affective Computing*, 2021.

[43] C. Wang, J. Zeng, S. Shan, and X. Chen, "Multi-task learning of emotion recognition and facial action unit detection with adaptively weights sharing network," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 56–60.

[44] S. Liu, E. Johns, and A. J. Davison, "End-to-end multi-task learning with attention," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1871–1880.

[45] X. Xiang and T. D. Tran, "Linear disentangled representation learning for facial actions," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 12, pp. 3539–3544, 2017.

[46] 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.

[47] W. Li, F. Abtahi, and Z. Zhu, "Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, July 2017, pp. 6766–6775.

[48] F. Benitez-Quiroz, Y. Wang, and A. Martinez, "Recognition of action units in the wild with deep nets and a new global-local loss," 10 2017, pp. 3990–3999.

[49] 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.

[50] S. R. Jannat, D. Fabiano, S. Canavan, and T. Neal, "Subject identification across large expression variations using 3d facial landmarks," *arXiv preprint arXiv:2005.08339*, 2020.

[51] Z. Zhao, Q. Liu, and F. Zhou, "Robust lightweight facial expression recognition network with label distribution training," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 4, 2021, pp. 3510–3519.

[52] B. Taha, M. Hayat, S. Berretti, D. Hatzinakos, and N. Werghi, "Learned 3d shape representations using fused geometrically augmented images: Application to facial expression and action unit detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 9, pp. 2900–2916, 2020.

[53] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3d face model for pose and illumination invariant face recognition," in *2009 sixth IEEE international conference on advanced video and signal based surveillance*. Ieee, 2009, pp. 296–301.

[54] J. Booth, A. Roussos, A. Ponniah, D. Dunaway, and S. Zafeiriou, "Large scale 3d morphable models," *International Journal of Computer Vision*, vol. 126, no. 2-4, pp. 233–254, 2018.

[55] L. Li, S. Wang, Z. Zhang, Y. Ding, Y. Zheng, X. Yu, and C. Fan, "Write-a-speaker: Text-based emotional and rhythmic talking-head generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, 2021, pp. 1911–1920.

[56] G. Yao, Y. Yuan, T. Shao, S. Li, S. Liu, Y. Liu, M. Wang, and K. Zhou, "One-shot face reenactment using appearance adaptive normalization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 4, 2021, pp. 3172–3180.

[57] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Nießner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt, "Deep video portraits," *TOG*, vol. 37, no. 4, pp. 1–14, 2018.

[58] Z. Geng, C. Cao, and S. Tulyakov, "3d guided fine-grained face manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9821–9830.

[59] W. Zhang, X. Ji, K. Chen, Y. Ding, and C. Fan, "Learning a facial expression embedding disentangled from identity," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6759–6768.

[60] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[61] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 67–74.

[62] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 9, pp. 1704–1716, 2011.

[63] J. Zhang, Y. Cao, and Q. Wu, "Vector of locally and adaptively aggregated descriptors for image feature representation," *Pattern Recognition*, vol. 116, p. 107952, 2021.

[64] H. Zhao, J. Jia, and V. Koltun, "Exploring self-attention for image recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 076–10 085.

[65] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[66] C. T. Kelley, *Iterative methods for optimization*. SIAM, 1999.

[67] J. Zhang, K. Chen, and J. Zheng, "Facial expression retargeting from human to avatar made easy," *IEEE Transactions on Visualization and Computer Graphics*, 2020.

[68] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, 2008.

[69] K. Zhao, W.-S. Chu, and H. Zhang, "Deep region and multi-label learning for facial action unit detection," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[70] 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*, 2019.

[71] 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, 2019, pp. 8594–8601.

[72] 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. 1, 2021.

[73] T. Song, Z. Cui, W. Zheng, and Q. Ji, "Hybrid message passing with performance-driven structures for facial action unit detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6267–6276.

[74] 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," *arXiv preprint arXiv:2205.01782*, 2022.

[75] 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.

[76] L. Zhong, Q. Liu, P. Yang, J. Huang, and D. Metaxas, "Learning multi-scale active facial patches for expression analysis," *IEEE Transactions on Cybernetics*, vol. 45, no. 8, pp. 1499–1510, 8 2015.
[77] 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.
[78] 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.
[79] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang *et al.*, "Multimodal spontaneous emotion corpus for human behavior analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3438–3446.
[80] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *FG 2018*. IEEE, 2018, pp. 59–66.
[81] 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.
[82] Z.-M. Chen, X.-S. Wei, P. Wang, and Y. Guo, "Multi-label image recognition with graph convolutional networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5177–5186.
[83] R. You, Z. Guo, L. Cui, X. Long, Y. Bao, and S. Wen, "Cross-modality attention with semantic graph embedding for multi-label classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 709–12 716.
[84] L. A. Jeni, J. F. Cohn, and F. De La Torre, "Facing imbalanced data–recommendations for the use of performance metrics," in *2013 Humaine association conference on affective computing and intelligent interaction*. IEEE, 2013, pp. 245–251.
[85] 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.
[86] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, A. Toisoul, B. W. Schuller *et al.*, "Sewa db: A rich database for audio-visual emotion and sentiment research in the wild," *IEEE transactions on pattern analysis and machine intelligence*, 2019.
[87] D. Kollias and S. Zafeiriou, "Aff-wild2: Extending the aff-wild database for affect recognition," *arXiv preprint arXiv:1811.07770*, 2018.
[88] J. Zhang, J. Yang, J. Yu, and J. Fan, "Semisupervised image classification by mutual learning of multiple self-supervised models," *International Journal of Intelligent Systems*, vol. 37, no. 5, pp. 3117–3141, 2022.
[89] J. Yu, Y. Rui, and D. Tao, "Click prediction for web image reranking using multimodal sparse coding," *IEEE Transactions on Image Processing*, vol. 23, no. 5, pp. 2019–2032, 2014.
[90] J. Yu, M. Tan, H. Zhang, Y. Rui, and D. Tao, "Hierarchical deep click feature prediction for fine-grained image recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 2, pp. 563–578, 2019.
[91] W. Zhang, F. Qiu, S. Wang, H. Zeng, Z. Zhang, R. An, B. Ma, and Y. Ding, "Transformer-based multimodal information fusion for facial expression analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2428–2437.

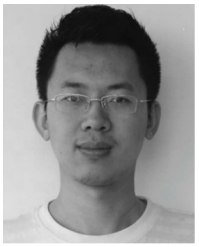
Wei Zhang received the B.E. degree in communication engineering from Nanjing University of Posts and Telecommunications, Jiangsu, China in 2017 and M.S. degree in electronic and information engineering from Zhejiang University, Zhejiang, China in 2020. She is currently a research scientist working with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

Lincheng Li received the B.S. and Ph.D. degree in electronic engineering from Tsinghua University, Beijing, China, in 2011 and 2017 respectively. He is currently a researcher in Netease Fuxi AI Lab, Hangzhou, China. His research interests include computer vision, pattern recognition, and image and video processing.

Yu Ding is currently an artificial intelligence expert and the leader of virtual human group at Netease Fuxi AI Lab, China. His research interests include virtual human, deep learning, image and video processing, talking-head generation, animation generation, multimodal computing, affective computing, nonverbal communication (face, gaze, and gesture), and embodied conversational agent. He received Ph.D. degree in Computer Science at Telecom Paris tech in Paris (France).

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.

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 or program co-chair for CASA 2014, SCA 2015, MIG 2022 and PG 223, he has been an Associate Editor for IEEE Transactions on Visualization and Computer Graphics, and Computer Graphics Forum. He is a distinguished member of ACM and a senior member of IEEE.



Xin Yu received the B.S. degree in electronic engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2009, the Ph.D. degree from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 2015, and the Ph.D. degree from the College of Engineering and Computer Science, Australian National University, Canberra, Australia, in 2019. He is currently a Senior Lecturer at the University of Queensland. His research interests include computer vision and image processing.