Local Relationship Learning with Person-specific Shape Regularization for Facial Action Unit Detection

Xuesong Niu[1,3], Hu Han[1,2], Songfan Yang[5,6], Yan Huang[6], Shiguang Shan[1,2,3,4]
[1] Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China
[2] Peng Cheng Laboratory, Shenzhen, China
[3] University of Chinese Academy of Sciences, Beijing 100049, China
[4] CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai, China
[5] College of Electronics and Information Engineering, Sichuan University, Chengdu, Sichuan, China
[6] TAL Education Group, Beijing, China
xuesong@vipl.ict.ac.cn, {hanhu, sgshan}@ict.ac.cn, {yangsongfan, galehuang}@100tal.com

Abstract

Encoding individual facial expressions via action units (AUs) coded by the Facial Action Coding System (FACS) has been found to be an effective approach in resolving the ambiguity issue among different expressions. While a number of methods have been proposed for AU detection, robust AU detection in the wild remains a challenging problem because of the diverse baseline AU intensities across individual subjects, and the weakness of appearance signal of AUs. To resolve these issues, in this work, we propose a novel AU detection method by utilizing local information and the relationship of individual local face regions. Through such a local relationship learning, we expect to utilize rich local information to improve the AU detection robustness against the potential perceptual inconsistency of individual local regions. In addition, considering the diversity in the baseline AU intensities of individual subjects, we further regularize local relationship learning via person-specific face shape information, i.e., reducing the influence of person-specific shape information, and obtaining more AU discriminative features. The proposed approach outperforms the state-of-the-art methods on two widely used AU detection datasets in the public domain (BP4D and DISFA).

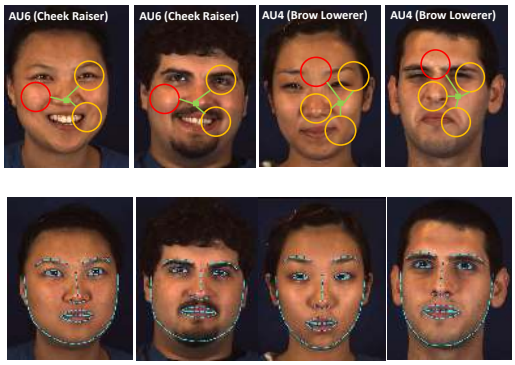

Figure 1: Each single local facial region defined for AUs in FACS (red circles) can be ambiguous because of face variations in pose, illumination, etc.; therefore, taking into account the relationship of multiple related face regions (yellow circles) can provide more robustness than using individual single local regions separately. At the same time, person-specific face shape information also influences the AU detection performance, i.e., detection of AU4 (Brow Lowerer) is highly influenced by the eye-eyebrow distance, which may vary significantly among different subjects. Therefore, we expect to reduce the influence of such person-specific shape information to the AU detection task, i.e., through regularization during feature learning.

1. Introduction

Facial expression is a natural and powerful means for human communications, which is highly associated with human's intention, attitude or mental state. Therefore, facial expression analysis has wide potential applications in diagnosing mental health [32], improving e-learning experiences [30], and detecting deception [12]. However, direct facial expression recognition in the wild can be challenging because of ambiguities between several expressions. One of the effective methods in resolving the ambiguity issue is to represent individual expression using the Facial Action Coding System (FACS) [10], in which each expression is identified as a specific configuration of multiple basic fa-

cial AUs. Therefore, a robust facial AU detection system is important for the accurate analysis of facial expressions.

Since different AUs correspond to different muscular activations of the face, the appearance of multiple local regions jointly reflects the presence of individual AUs, and the local information is crucial for AU detection. The early works on facial AUs detection represented different local facial areas using the traditional hand-crafted features, which can be not discriminative enough for capturing the facial morphology [46, 48]. Recently, deep learning has been widely applied for facial representation learning, including using the deep representation for more effective AU detection [5, 8, 24, 35, 47].

However, besides learning more AU discriminative features, the relationship of individual facial regions can be very important for AU detection. As shown in Fig. 1, each single local face region defined in FACS can be ambiguous for AU detection because of face variations in pose, illumination, etc.; therefore, taking into account the relationship of multiple face regions can provide more robustness than using a single local region. For instance, the cheek area and the mouth corner of the face usually active simultaneously in a common facial behavior called Duchenne smile, resulting in high correlations between AU6 (cheek raiser) and AU12 (lip corner puller). Some approaches tried to utilize such local relationship information by using multi-label learning [24, 46, 47], but only holistic feature representations were used. A meticulous modeling method is required for effectively leveraging the relationship of different local facial regions to perform robust AU detection.

Another critical characteristic of AU is that the appearance of the same AU may vary among different subjects due to the different morphological aspects and ways to express the emotions of different subjects (see Fig. 1). This is the reason why designing a person-specific AU detector can improve the AU detection accuracy. However, existing person-specific AU detection methods require either retraining the model for the new subjects [7, 43], or additional data of the new subjects for model generation [1, 33] or normalization [2]. These constraints limit the range of applications of the existing AU detection methods.

In this paper, we propose an end-to-end trainable network for AU detection using **L**ocal relationship learning with **P**erson-specific shape regularization (namely LP-Net). The LP-Net consists of a stem network, a local relationship learning module (L-Net) and a person-specific shape regularization module (P-Net). The stem network mainly contains convolutional layers for local region feature extraction. The extracted local features are then fed to the local relationship learning module for relationship learning and predicting the AU occurrence probabilities. At the same time, P-Net aims to learn features that are independent with the features by L-Net, and thus works as a regularization term to reduces the influence of person-specific shape information. As a result, the final features learned by L-Net are more discriminative and generalizable for AU detection.

The contributions of this work are three-fold: (i) we propose a novel end-to-end trainable framework for AU detection, which is able to leverage not only the local information but also the relationship of individual regions to improve the AU detection robustness; (ii) we regularize local relationship learning via person-specific face shape information to obtain more discriminative and generalizable features related to AU detection; (iii) the proposed approach outperforms the state-of-the-art methods on two widely used AU detection datasets BP4D and DISFA.

2. Related Work

Automatic facial action unit detection has been studied for decades, and several works have been proposed. Various features [4, 20, 25, 26] and classifiers [7, 38, 44, 46] have been applied to build a robust facial action unit detection system under realistic situations. Recently, CNNs have shown great power in many computer vision tasks such as face verification [37], objection detection [13], and image recognition [17], and have been successfully applied to automatic facial action unit detection [5, 15]. The reader can refer the recent surveys and challenges [9, 27, 39] for more information. In the following paragraphs, we will review the relative works to ours.

Since facial AUs are defined as patterns of different facial muscular movements, the ways they perform the facial expressions are relatively based on the local facial appearance. Several works are based on this character and use local information for facial AU detection. Zhong *et al*. [48] divided the face area into multiple uniform patches, and use the common and specific patches to describe different expressions. Taheri *et al*. [36] defined fixed regions for different AUs and used sparse coding to recover facial expressions using the composition rules of AUs. Zhao *et al*. [46] performed a patch selection method based on facial landmarks and group sparsity learning. All these methods used traditional features to represent the face local information and these features are not sufficiently expressive.

Besides of traditional features, the great modeling power of CNNs has also been successfully leveraged to facial action unit detection. In [47], Zhao *et al*. proposed a region layer to induce the CNN to focus on important facial regions for better feature learning. In [23], Li *et al*. trained different CNNs using different parts of a face and merged the features from different areas in an early fusion fashion using fully connected layers. In [24], Li *et al*. proposed a local feature learning method based on enhanced and cropped facial area. In [35], Shao *et al*. proposed an end-to-end deep learning framework for joint AU detection and face alignment, which used the alignment feature to compute

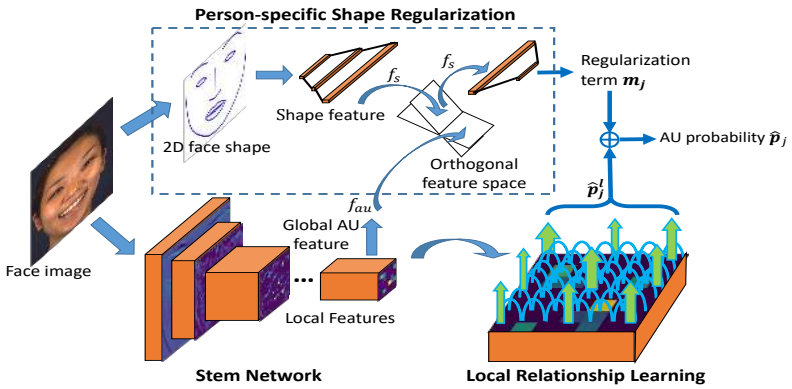

Figure 2: An overview of our approach for AU detection, which consists of a stem network, a local relationship learning module (L-Net) and a person-specific shape regularization module (P-Net). By using P-Net to model the person-specific shape information, and enforcing the person-specific shape features are independent with the features learned by L-Net, we expect the final features for AU detection can be more discriminative and generalizable.

an adaptive attention map for better local feature learning. These methods have drastically improved the performance of facial AU detection with the great modeling power of deep learning. However, all these methods only focused on different regions and failed to consider the relationship of different local areas. At the same time, the appearance of different facial local areas usually changes simultaneously because of the underlying facial anatomy, and this relationship of different local regions will also benefit the detection of AUs.

Besides directly using local features to predict AUs, another way of modeling the relationship of AUs is to use the correlations of different AUs. Walecki *et al*. [40] proposed a method to model the AU relations and feature representations simultaneously by combining conditional random field (CRF) with deep learning. In [41], Wang *et al*. proposed a restricted Boltzmann machine to capture high-order AU interactions. In [8], Corneanu *et al*. applied a graphical model inference approach to passing AU probabilities between different AU labels. All these methods took the probabilistic dependencies between different AUs into consideration and used the correlations to refine the predicted results. However, most of these methods computed the AU probabilities using the feature generated from the entire face area. Local information has been ignored, which can be very important for facial AU detection. At the same time, AU-to-AU relationships are mainly generated using facial anatomy and FACS [10] based on posed expressions, and their generalization ability to spontaneous expressions is not known.

Another key characteristic of AUs is that the appearance of the same AU may vary among different subjects. This is the reason why many person-specific AU analysis models have been proposed, and have been found to be effective for AU detection. Chu *et al*. [7] proposed a selective transfer machine to personalize the AU detector by re-weighting of the source distribution to match that of the target distribution. Zeng *et al*. [43] applied a similar re-weighting strategy and learned a person-specific classifier using synthetic labels provided by confident classifiers. This kind of person-specific AU detectors requires re-training the model for each subject, which can be time-consuming. Besides re-weighting the source distribution, Sangineto *et al*. [33] proposed a transfer process to learn discriminative mappings between the data distribution associated with each source subject and the corresponding parameters. Almaev *et al*. [1] proposed a multi-task learning structure to learn the latent relations among tasks using one single AU and transfer the latent relations to other AUs. In [2], Baltrušaitis *et al*. proposed a simple but efficient way for person-specific feature normalization using the median of all the features in a video. Although all these methods do not need to re-train the model for a new subject, they still need additional data to generate a new AU predictor, which limits the application scope in practical scenarios.

In contrast to these existing methods, we employ an end-to-end deep framework LP-Net to predict AUs. We not only consider the local information for facial AUs prediction but also take the relationship of different facial regions into consideration. At the same time, person-specific shape regularization is also utilized to reduce the influence of the diverse baseline AU intensities among different subjects.

3. Proposed Method

Fig. 2 shows the overall framework of our LP-Net for facial AU detection, which consists of a stem network, a local relationship learning module (L-Net) and a person-specific shape regularization module (P-Net). We detail the proposed approach in the following sections.

3.1. Overview of LP-Net

Feature representation is the key component of building a robust AU detection system, in which CNN has shown its great power and achieved great success in many computer vision tasks [13, 17, 37]. Traditional CNNs usually feed the output of convolutional layers to a global pooling layer in order to get a robust global feature. However, such an operation would fail to capture the local information for structured objects like faces and thus ignoring some local but important information related to AU detection.

To overcome these limitations, as shown in Fig. 2, we remove the global pooling layer in CNN and directly use the output feature maps from the convolutional layers as the representation of local features. CNN networks like ResNet [17] have been proved to have a strong ability for local features generation with only convolutional layers. So, here, we choose ResNet-34 [17] as our stem network for local feature learning. The output of the last convolutional layer of ResNet-34, which contains 512 feature maps with a size of 7×7, is regarded as the set of local features and utilized for further processing. Thus, we in total obtain 49 local features of 512-dimension from the stem network.

After we get the local features generated from the stem network, a local relationship learning module based on Long Short-Term Memory (LSTM) [18] (L-Net) is introduced to automatically explore the underline relationship of individual local facial regions in the feature space. Our L-Net jointly considers the features of local regions and their relationship and outputs the probabilities of individual AUs.

As summarized in Section 1, another challenge is that different subjects may have different baseline AU intensities because of the face shape differences. A person-specific shape regularization module (P-Net) is used to model such person-specific information based on 2D face shape. The features encoded by P-Net are expected to be independent with the features encoded by L-Net, and further used to calculate the regularization term to refine the AU probabilities predicted by L-Net. Thus P-Net works as a regularization module to enforce the L-Net to learn more subject-independent features for AU detection, and the refined AU probabilities by P-Net are used as the final prediction of our LP-Net.

3.2. Local Relationship Learning via L-Net

Fig. 3 gives the detailed structure of our L-Net for local relationship learning. Since the feature maps generated by the stem network are from the last convolutional layer of ResNet-34, each element ($1 \times 1 \times 512$) in the feature map highlights the characteristic of a facial region. Therefore, we use each element on the feature maps as a representation of the local face area and use it to perform local relationship learning.

Specifically, we get k local features $f_1, f_2, \cdots, f_k$ from

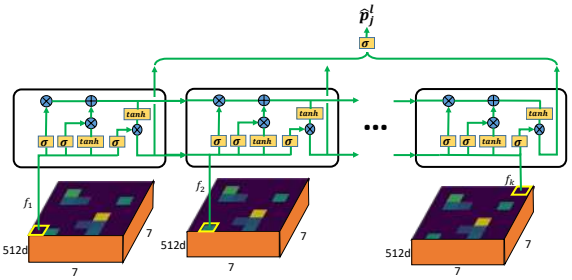

Figure 3: Each element (49 elements in total) of the feature map generated by the stem network is treated as a representation of a local region and used as the input of our L-Net based on LSTM. L-Net explores the underline relationship of individual local regions, and outputs the probabilities.

the stem network ($k = 49$ for ResNet-34). Each local feature f_i will be used for AU prediction, and output an AU occurrence probability. The LSTM structure is utilized to learn the relationship and outputs the probabilities of different local features. Since different AUs have different muscular activations, and the contributions of individual local features for predicting the probability should be different. Therefore, we predict the occurrence probability of each AU separately, i.e., using C LSTM structures to predict the probabilities of all the C AUs.

At the same time, we believe that every local feature can be helpful for detecting individual AUs, and thus all the k local features are fed to each LSTM structure. The final decision for the detection of each AU is obtained by combining all prediction results and the final predicted AU occurrence probabilities by L-Net can be written as

$$\hat{p_j^l} = \sigma(\frac{1}{k}\sum_{i=1}^{k} LSTM_j(f_i)) \quad j = 1, 2, \cdots, C \tag{1}$$

where σ is a sigmoid function.

3.3. Person-specific Shape Regularization via P-Net

P-Net aims to reduce the influence of person-specific shape information and obtaining more discriminative and general features for AU detection. As shown in Fig. 4, we use 2D facial landmarks as a representation of the face shape [14, 21]. Specifically, we use a robust facial landmarks detector (Convolutional Experts Constrained Local Model [3, 42]) to detect 68 facial landmarks $P_1, P_2, \cdots, P_{68}$, and then all the face images based on the two eye centers to reduce the influence of head pose. After the face images are aligned, each landmark point is normalized using

$$P_{norm} = \frac{P - P_{center}}{d} \tag{2}$$

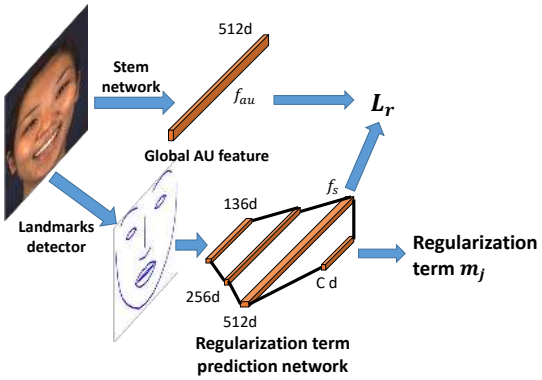

Figure 4: The detailed structure of our person-specific shape regularization network (P-Net). The 68 facial landmarks are treated as the representation of face shape, and used for regularization terms calculation. A regularization loss L_r is applied to guide the P-Net to output the AU-independent person-specific facial shape regularization term m_j for C AUs, which are further used for refining the AU occurrence probabilities predicted by L-Net.

where P_{center} is the center point of the two eyes, and d is the interpupillary distance (IPD).

The normalized landmarks are used as the input to our P-Net in order to predict the person-specific shape regularization term (see Fig. 4). We expect the P-Net only learn the AU-independent person-specific face shape information, so we propose a regularization loss L_r aiming for orthogonalizing the features learned by P-Net and the features used for AU detection by L-Net. The loss is formulated as

$$L_r = |f_{au} \bullet f_s| \tag{3}$$

where $\bullet$ represents the inner product of two vectors, f_{au} is the average of the k local features generated from the stem network, and f_s is the last layer feature of P-Net for regularization term prediction. For each input image, we calculate the regularization terms $m_1, m_2, \cdots, m_c$ for all the C AUs and use them to refine the predicted probabilities by L-Net. The final predicted probability $\hat{p_1}, \hat{p_2}, \cdots, \hat{p_c}$ of the LP-Net for all the C AUs can be written as

$$\begin{aligned} \hat{p_j} &= \sigma(\frac{1}{k}\sum_{i=1}^{k} LSTM_j(f_i) + m_j) \\ j &= 1, 2, \cdots, C \end{aligned} \tag{4}$$

AU prediction is a multi-label binary classification problem, and for most of the AU prediction benchmarks, the occurrences of AUs are highly imbalanced [9, 27, 39]. To better handle such a multi-label and imbalance problem, we choose to use binary cross entropy loss L_{au} with Selective Learning [16] as our loss function

$$L_{au} = -\frac{1}{C}\sum_{j=1}^{C} w_c[p_j \log \hat{p_j} + (1-p_j)\log(1-\hat{p_j})] \tag{5}$$

where p_j represents the ground-truth probability of the occurrence for the j-th AU, with 1 denoting occurrence of an AU and 0 denoting no occurrence. $\hat{p_j}$ is the predicted probability by our LP-Net. The weight w_c is a balancing parameter which is calculated in each batch using the Selective Learning strategy [16]. The overall loss function of the proposed LP-Net can be written as

$$L_{all} = L_{au} + \lambda L_r \tag{6}$$

where λ is a hyper-parameter that balances the influences of the two losses.

4. Experimental Results

In this section, we provide experimental evaluations on several public-domain AU detection databases and give detailed analysis of the experimental results.

4.1. Experimental Settings

4.1.1 Database

We evaluate our LP-Net on two spontaneous databases BP4D [45] and DISFA [28], which have been widely used for facial AUs detection. **BP4D** is a spontaneous facial expression database containing 328 videos for 41 participants (23 females and 18 males). Each subject is involved in 8 sessions, and their spontaneous facial actions are captured with both 2D and 3D videos. 12 AUs are coded for the 328 videos, and there are about 140,000 frames with AU labels of occurrence or absence. **DISFA** consists of 27 videos from 12 females and 15 males. Each subject is asked to watch a 4-minute video to elicit facial AUs. 12 AUs are labeled with AU intensity from 0 to 5 for each video. About 130,000 frames are used in the final experiments. Following the experiment setting of [8, 23, 24, 35], we conduct a subject-exclusive 3-fold cross-validation on BP4D, and further fine-tune the best model trained on BP4D for AU detection on DISFA under a subject-exclusive 3-fold validation protocol. For DISFA databse, 8 of the 12 AUs are used for evaluation and the frames with AU intensity equal or greater than 2 are selected as positive samples and the rest are selected as negative samples.

4.1.2 Image Pre-processing

For each input image, the CE-CLM facial landmark detector is used to estimate the 68 facial landmarks (see Fig. 4). Then following the idea of Baltrušaitis *et al*. [2], all the faces are aligned and masked using a similarity transform based on the detected landmarks to reduce the variations of pose and scale. All the aligned face images are resized to 240×240 and then randomly cropped to 224×224 for training. Images center-cropped from the aligned faces are utilized for testing. We also use random horizontal flip and random rotation for data augmentation.

4.1.3 Training

We incrementally train each part of our LP-Net. First, we pre-train our stem network on the face recognition database VGGFace2 [6]. Then we train the stem network on the AU databases. An Adam optimizer with an initial learning rate of 0.001 is applied for optimizing the stem network. After that, we add the L-Net module and jointly train the stem network and the L-Net. The initial learning rate is set to 0.0005 for the stem network and 0.001 for L-Net. Next, we add the P-Net to the network and jointly train the whole network with an initial learning rate of 0.0005 for stem network and L-Net and an initial learning rate of 0.001 for P-Net. The max iteration for all the training steps is set to 30 epochs, and the batch size is set to 100. The balance parameter λ for regularization loss L_r is set to 1. All the implementations are based on PyTorch [31].

4.1.4 Evaluation Metrics

We evaluate the performance of all methods using F1-frame score [19]. F1-frame score is the harmonic mean of precision and recall of frame-based AU detection and has been widely used for AU detection. For each method, F1-frame for all the AUs are calculated and then averaged (donated as **Avg.**) for evaluation.

4.2. Results

4.2.1 Comparisons with the State-of-the-art

We first compare our LP-Net against the state-of-the-art methods under the same subject-exclusive three-fold cross-validation protocol. Traditional methods LSVM [11], JPML [46], APL [48], and CPM [43], and deep learning methods DRML [47], EAC-Net [24], ROI [23], DSIN [8], and JAA-Net [35] are used for comparison. Since we focus on image-based AU detection in this work, the video-based methods such as ROI-LSTM [23] are not used for comparison. At the same time, we notice some methods such as DSIN [8] used threshold turning per AU, while most of the other baseline methods did not use threshold turning per AU. So for fair comparisons, we report the performance of individual methods without threshold tuning per AU. For the baseline methods LSVM [11], JPML [46], APL [48], and CPM [43], we directly use their results reported in [24, 35, 47].

Table 1 shows the results of different methods on the BP4D database. It can be seen that our LP-Net outperforms all the baseline approaches on this challenging spontaneous facial expression database. Comparing LP-Net with the state-of-the-art methods based on deeply-learned local features such as ROI [23], DRML [47], JAA-Net [35] and DSIN [8], our LP-Net could achieve the best or second-best detection performance for most of the 12 AUs annotated in BP4D. We also achieve the best performance in terms of average F1-frame score. At the same time. Our LP-Net also outperforms the person-specific AU detection models, such as CPM [43], by a large margin, which indicates that our P-Net is very effective in dealing with the challenge of diverse baseline AU intensities among different subjects.

When comparing with the state-of-the-art methods [35, 8], we also find that the performance of our LP-Net drops when the facial regions of the AUs are small, such as AU1 and AU2. The reason is that the local features are generated from the last layer of the Stem-Net, which are high-level in semantics and may be not sensitive in representing small regions. However, although the performance drops when directly using the Stem-Net for local features generation, the computational complexity is significantly reduced because our LP-Net does not need an additional backbone networks [8] for local feature generation or an additional branch to enhance the local feature [35].

Experimental results on the DISFA database are reported in Table 2. It can be observed that our LP-Net again outperforms all the state-of-the-art methods. We achieve the best performance on most AUs, as well as the average F1-frame score for all AUs. These results suggest that our LP-Net has a good generalization ability.

4.2.2 Ablation Study

We provide ablation study to investigate the effectiveness of each part of our LP-Net. Table 3 shows the F1-frame scores for each AU as well as the average F1-frame score by individual ablation experiments on BP4D.

Choice of Stem Network: In our LP-Net, stem network is used for local features generation. We choose ResNet as the stem network and three commonly used networks (RestNet-18, ResNet-34, and Rest-Net50) have been considered. The results are shown in Table 3. From the results, we can see that ResNet-34 outperforms ResNet-18 with an improvement of average F1-frame from 52.9 to 53.7, indicating that the deeper network could give richer features for AU detection. However, when the network is further deepened to ResNet-50, the performance drops to 52.5. The possible reason is that the AU databases have limited subjects and a very deep network may suffer from over-fitting. We use ResNet-34 in our following experiments.

Data Balancing with Selective Learning: Since it is complicated to collect and annotate AUs for a large face database, most AU databases are highly imbalanced. After we apply the Selective Learning strategy [16] for data balancing, the average F1-frame on BP4D has been improved from 53.7 to 55.2, indicating the effectiveness of Selective Learning [16] used in our LP-Net.

Data Augmentation and Model Pre-training: Because of the difficulties of AU data collection, there are usually limited subjects in AU databases. Data augmentation and model

Table 1: F1-frame score (in %) for 12 AUs reported by the proposed LP-Net and the state-of-the-art methods on the BP4D database. The best and second are indicated using brackets and bold, and brackets alone, respectively.

Method	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU14	AU15	AU17	AU23	AU24	**Avg.**
LSVM [11]	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 [46]	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 [47]	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
CPM [43]	43.4	40.7	43.3	59.2	61.3	62.1	68.5	52.5	36.7	54.3	39.5	37.8	50.0
EAC-Net [24]	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
ROI [23]	36.2	31.6	43.4	[77.1]	73.7	**[85.0]**	[87.0]	62.6	**[45.7]**	58.0	38.3	37.4	56.4
JAA-Net [35]	[47.2]	**[44.0]**	[54.9]	**[77.5]**	[74.6]	[84.0]	86.9	61.9	43.6	60.3	[42.7]	41.9	[60.0]
DSIN [8]	**[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
LP-Net	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]**

Table 2: F1-frame score (in %) for 8 AUs reported by the proposed LP-Net and the state-of-the-art methods on the DISFA database. The best and second are indicated using brackets and bold, and brackets alone, respectively.

Method	AU1	AU2	AU4	AU6	AU9	AU12	AU25	AU26	**Avg.**
LSVM [11]	10.8	10.0	21.8	15.7	11.5	70.4	12.0	22.1	21.8
DRML [47]	17.3	17.7	37.4	29.0	10.7	37.7	38.5	20.1	26.7
APL [48]	11.4	12.0	30.1	12.4	10.1	65.9	21.4	26.9	23.8
ROI [24]	41.5	26.4	66.4	**[50.7]**	8.5	89.3	88.9	15.6	48.5
JAA-Net [35]	**[43.7]**	**[46.2]**	56.0	41.4	44.7	69.6	88.3	[58.4]	[56.0]
DSIN [8]	[42.4]	[39.0]	[68.4]	28.6	[46.8]	[70.8]	[90.4]	42.2	53.6
LP-Net	29.9	24.7	**[72.7]**	[46.8]	**[49.6]**	**[72.9]**	**[93.8]**	**[65.0]**	**[56.9]**

pre-training are commonly used strategies to reduce the risk of modeling overfitting. With data augmentation, the average F1-frame on BP4D has been improved from 55.2 to 56.5 and further improved to 58.0 when pre-training the model on VGGFace2 [6]. The results indicate that data augmentation and model pre-training are effective ways to improve AU detection performance.

Effectiveness of Local Relationship Learning: In order to illustrate the effectiveness of local information and local relationship learning, we first conduct the experiments by predicting AU probabilities from every local feature and fusing all the results with mean pooling. This baseline method is donated as *Stem-Net+LF*, and it achieves a better average F1-frame of 58.8 than Stem-Net (see Table 3). The results indicate that local features are more representative for AU detection. However, directly fusing all the AU probabilities predicted by local features with mean pooling ignores the local relationship of different local regions. We further add the local relationship learning module to the Stem-Net (donated as *Stem-Net+L-Net* in Table 3). The average F1-frame is improved from 58.0 to 60.2, indicating that the relationship of different local regions are useful for AU predictions and our L-Net is effective for modeling this kind of information. Fig. 5 shows some example class activation maps [34] for AU4 and AU23, and we can see that by using the proposed L-Net, the model can focus more on the related areas of the concerned AUs.

We further conduct experiment using all the local features in the activation areas of different AUs defined in FACS as the input of our relationship learning module to see whether all the features are useful for AU detection. The network is donated as *Stem-Net+FACS* in Table 3. From the results, we can see that with relationship learning, *Stem-Net+FACS* outperforms the network that only uses local features for AU prediction (*Stem-Net+LF*), and achieves an average F1-frame score of 59.0. This again shows that the local relationship is useful for improving the AU detection performance. At the same time, when we take all the local information into consideration, we could achieve a better result. This indicates that using all the local features is helpful for AU detection. The possible reason is that the local features are generated from the deep Stem-Net, which contains rich information for AU detection.

Effectiveness of Person-specific Shape Regularization: In order to illustrate the effectiveness of the person-specific shape regularization module (P-Net), we conduct the experiments with and without using the regularization loss L_r. When the regularization loss is not used, the average F1-frame can be improved from 58.0 to 58.7 because the face shape information is added for AU detection. If we add the regularization loss to decompose the AU features and face shape features, the average F1-frame will be improved to

Table 3: F1-frame score (in %) of ablation experiments for 12 AUs on the BP4D database. The best and second are indicated using brackets and bold, and brackets alone, respectively.

Method	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU14	AU15	AU17	AU23	AU24	**Avg.**
ResNet-18	42.3	30.3	37.1	74.1	72.4	81.9	83.6	57.1	34.2	54.5	37.0	31.1	52.9
ResNet-34	38.3	31.7	44.0	73.5	71.7	80.2	84.8	58.6	32.5	54.6	35.8	39.0	53.7
ResNet-50	38.2	31.1	40.3	73.6	69.7	80.5	82.9	55.1	29.1	56.2	37.1	36.0	52.5
ResNet-34+SL	38.6	34.0	46.9	72.0	73.0	79.8	84.5	60.7	38.7	60.0	33.0	41.5	55.2
ResNet-34+SL/DA	41.7	31.0	47.9	75.2	76.9	80.0	85.5	60.3	35.9	58.5	37.6	47.8	56.5
ResNet-34+SL/DA/P	41.3	37.7	49.8	[**77.1**]	75.6	81.8	86.4	61.7	41.1	58.1	42.4	43.5	58.0
Stem-Net*+LF	40.5	36.2	48.0	76.2	77.4	82.6	86.0	62.8	42.8	60.3	[46.0]	46.7	58.8
Stem-Net*+FACS	41.0	[**39.9**]	52.0	74.7	75.1	81.5	85.4	[63.1]	44.7	58.7	41.7	50.3	59.0
Stem-Net*+L-Net	41.2	[39.7]	50.8	[76.3]	77.9	81.7	86.2	61.7	[46.2]	[62.1]	45.5	53.2	[60.2]
Stem-Net*+P-Net w/o L_r	42.3	36.1	50.3	75.8	[78.3]	82.4	[86.8]	61.5	[**46.6**]	60.7	42.5	41.5	58.7
Stem-Net*+P-Net	[**43.5**]	35.8	[53.1]	73.8	[**78.5**]	[83.1]	85.6	58.5	43.0	[**62.4**]	43.7	[**56.3**]	59.8
LP-Net	[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**]

SL: Selective Learning balancing; **DA**: Data augmentation; **P**: Pre-training model on VGGFace2;
LF: AU detection with all local features; **FACS**: AU detection with local features in areas defined by FACS;
* Stem-Net represents ResNet-34 + SL/DA/P.

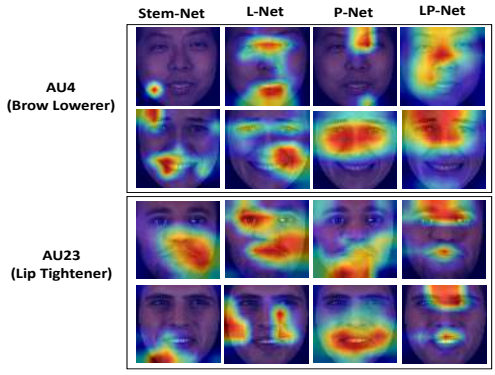

Figure 5: Class activation maps [34] that show the discriminative regions for AU4 and AU23. The class activation maps for Stem-Net, L-Net, P-Net and LP-Net are listed from left to right. The first two rows are activation maps for AU4 and the bottom two rows are for AU23.

59.8, indicating that the regularization terms predicted by P-Net are useful in reflecting the diverse baseline AU intensities. The class activation maps [34] using P-Net are also shown in Fig. 5. From the activation maps, we can see that the network is more likely to focus on the informative regions with P-Net.

When both local relationship learning module and person-specific shape regularization module are included to the Stem-Net (LP-Net), the network can focus on both the concerned AU regions and the related facial regions (see Fig. 5), and thus is able to achieve a better performance (an average F1-frame of 61.0).

5. Conclusion

Robust facial action unit (AU) detection in the wild remains a challenging problem due to the diversity of expression intensities across individual subjects and variation of facial appearance due to pose, illumination, etc. While the Facial Action Coding System (FACS) has been proven to be an effective approach in resolving ambiguity in AU detection, the information of local face regions and their relationship are still not fully exploited to achieve robust AU detection. We propose a novel end-to-end trainable framework (LP-Net) for AU detection, which consists of three modules (Stem-Net, L-net, and P-Net) for share feature learning, local relationship modeling, and person-specific shape regularization, respectively. The proposed approach outperforms the state-of-the-art methods on two widely used AU detection datasets in the public domain.

In our future work, we would like to explore different approaches for modeling the local relationship, *e.g.*, through conditional random field [22], graph convolutional networks [29], etc. In addition, learning features covering diverse scales will also be taken into consideration.

Acknowledgment

This research was supported in part by the Natural Science Foundation of China (grants 61732004 and 61672496), External Cooperation Program of Chinese Academy of Sciences (CAS) (grant GJHZ1843), and Youth Innovation Promotion Association CAS (grant 2018135).

References

[1] Timur Almaev, Brais Martinez, and Michel Valstar. Learning to transfer: transferring latent task structures and its application to person-specific facial action unit detection. In *Proc. IEEE ICCV*, pages 3774–3782, 2015. 2, 3

[2] Tadas Baltrušaitis, Marwa Mahmoud, and Peter Robinson. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *Proc. IEEE FG*, pages 1–6, 2015. 2, 3, 5

[3] Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency. Constrained local neural fields for robust facial landmark detection in the wild. In *Proc. IEEE ICCV Workshops*, pages 354–361, 2013. 4

[4] Marian Stewart Bartlett, Gwen Littlewort, Mark G Frank, Claudia Lainscsek, Ian R Fasel, Javier R Movellan, et al. Automatic recognition of facial actions in spontaneous expressions. *J Multimed.*, 1(6):22–35, 2006. 2

[5] Carlos Fabian Benitez-Quiroz, Yan Wang, and Aleix M Martinez. Recognition of action units in the wild with deep nets and a new global-local loss. In *Proc. IEEE ICCV*, pages 3990–3999, 2017. 2

[6] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *Proc. IEEE FG*, pages 67–74, 2018. 6, 7

[7] Wen-Sheng Chu, Fernando De la Torre, and Jeffery F Cohn. Selective transfer machine for personalized facial action unit detection. In *Proc. IEEE CVPR*, pages 3515–3522, 2013. 2, 3

[8] Ciprian Corneanu, Meysam Madadi, and Sergio Escalera. Deep structure inference network for facial action unit recognition. In *Proc. ECCV*, pages 298–313, 2018. 2, 3, 5, 6, 7

[9] Abhinav Dhall, Amanjot Kaur, Roland Goecke, and Tom Gedeon. Emotiw 2018: Audio-video, student engagement and group-level affect prediction. In *Proc. ICMI*, pages 653–656, 2018. 2, 5

[10] Paul Ekman and Erika L Rosenberg. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997. 1, 3

[11] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9(Aug):1871–1874, 2008. 6, 7

[12] Robert S Feldman, Larry Jenkins, and Oladeji Popoola. Detection of deception in adults and children via facial expressions. *Child development*, pages 350–355, 1979. 1

[13] Ross Girshick. Fast R-CNN. In *Proc. IEEE ICCV*, pages 1440–1448, 2015. 2, 4

[14] Hu Han, Brendan F Klare, Kathryn Bonnen, and Anil K Jain. Matching composite sketches to face photos: A component-based approach. *IEEE Trans. Inf. Forensics Security*, 8(1):191–204, 2013. 4

[15] Shizhong Han, Zibo Meng, Ahmed-Shehab Khan, and Yan Tong. Incremental boosting convolutional neural network for facial action unit recognition. In *Proc. NeurIPS*, pages 109–117, 2016. 2

[16] Emily M Hand, Carlos D Castillo, and Rama Chellappa. Doing the best we can with what we have: Multi-label balancing with selective learning for attribute prediction. In *Proc. AAAI*, 2018. 5, 6

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE CVPR*, pages 770–778, 2016. 2, 4

[18] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997. 4

[19] László A Jeni, Jeffrey F Cohn, and Fernando De La Torre. Facing imbalanced data–recommendations for the use of performance metrics. In *Proc. IEEE ACII*, pages 245–251, 2013. 6

[20] Bihan Jiang, Michel F Valstar, and Maja Pantic. Action unit detection using sparse appearance descriptors in space-time video volumes. In *Proc. IEEE FG*, pages 314–321, 2011. 2

[21] Scott J Klum, Hu Han, Brendan F Klare, and Anil K Jain. The facesketchid system: Matching facial composites to mugshots. *IEEE Trans. Inf. Forensics Security*, 9(12):2248–2263, 2014. 4

[22] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. ICML*, 2001. 8

[23] Wei Li, Farnaz Abtahi, and Zhigang Zhu. Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In *Proc. IEEE CVPR*, pages 6766–6775, 2017. 2, 5, 6, 7

[24] Wei Li, Farnaz Abtahi, Zhigang Zhu, and Lijun Yin. EAC-Net: Deep nets with enhancing and cropping for facial action unit detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(11):2583–2596, 2018. 2, 5, 6, 7

[25] Gwen Littlewort, Marian Stewart Bartlett, Ian Fasel, Joshua Susskind, and Javier Movellan. Dynamics of facial expression extracted automatically from video. *Image Vis. Comput.*, (6):615–625, 2006. 2

[26] Simon Lucey, Ahmed Bilal Ashraf, and Jeffrey F Cohn. Investigating spontaneous facial action recognition through AAM representations of the face. In *Face recognition*. IntechOpen, 2007. 2

[27] Brais Martinez, Michel F Valstar, Bihan Jiang, and Maja Pantic. Automatic analysis of facial actions: A survey. *IEEE Trans. Affect. Comput.*, 2017. 2, 5

[28] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. DISFA: A spontaneous facial action intensity database. *IEEE Trans. Affect. Comput.*, 4(2):151–160, 2013. 5

[29] Mathias Niepert, Mohamed Ahmed, and Konstantin Kutzkov. Learning convolutional neural networks for graphs. In *Proc. ICML*, pages 2014–2023, 2016. 8

[30] Xuesong Niu, Hu Han, Jiabei Zeng, Xuran Sun, Shiguang Shan, Yan Huang, Songfan Yang, and Xilin Chen. Automatic engagement prediction with GAP feature. In *Proc. ICMI*, pages 599–603, 2018. 1

[31] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban

Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *Proc. NeurIPS Workshops*, 2017. 6

[32] David R Rubinow and Robert M Post. Impaired recognition of affect in facial expression in depressed patients. *Biol. Psychiatry*, 31(9):947–953, 1992. 1

[33] Enver Sangineto, Gloria Zen, Elisa Ricci, and Nicu Sebe. We are not all equal: Personalizing models for facial expression analysis with transductive parameter transfer. In *Proc. ACM MM*, pages 357–366, 2014. 2, 3

[34] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proc. ICCV*, pages 618–626, 2017. 7, 8

[35] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. Deep adaptive attention for joint facial action unit detection and face alignment. In *Proc. ECCV*, pages 705–720, 2018. 2, 5, 6, 7

[36] Sima Taheri, Qiang Qiu, and Rama Chellappa. Structure-preserving sparse decomposition for facial expression analysis. *IEEE Trans. Image Process.*, 23(8):3590–3603, 2014. 2

[37] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proc. IEEE CVPR*, pages 1701–1708, 2014. 2, 4

[38] Yan Tong and Qiang Ji. Learning bayesian networks with qualitative constraints. In *Proc. IEEE CVPR*, pages 1–8. IEEE, 2008. 2

[39] Michel F Valstar, Enrique Sánchez-Lozano, Jeffrey F Cohn, László A Jeni, Jeffrey M Girard, Zheng Zhang, Lijun Yin, and Maja Pantic. FERA 2017-addressing head pose in the third facial expression recognition and analysis challenge. In *Proc. IEEE FG*, pages 839–847, 2017. 2, 5

[40] Robert Walecki, Ognjen Rudovic, Vladimir Pavlovic, Björn Schuller, and Maja Pantic. Deep structured learning for facial action unit intensity estimation. In *Proc. IEEE CVPR*, pages 5709–5718, 2017. 3

[41] Ziheng Wang, Yongqiang Li, Shangfei Wang, and Qiang Ji. Capturing global semantic relationships for facial action unit recognition. In *Proc. IEEE ICCV*, pages 3304–3311, 2013. 3

[42] Amir Zadeh, Yao Chong Lim, Tadas Baltrusaitis, and Louis-Philippe Morency. Convolutional experts constrained local model for 3D facial landmark detection. In *Proc. IEEE ICCV Workshops*, pages 2519–2528, 2017. 4

[43] Jiabei Zeng, Wen-Sheng Chu, Fernando De la Torre, Jeffrey F Cohn, and Zhang Xiong. Confidence preserving machine for facial action unit detection. In *Proc. IEEE ICCV*, pages 3622–3630, 2015. 2, 3, 6, 7

[44] Xiao Zhang and Mohammad H Mahoor. Task-dependent multi-task multiple kernel learning for facial action unit detection. *Pattern Recognit*, 51:187–196, 2016. 2

[45] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. BP4D-spontaneous: a high-resolution spontaneous 3D dynamic facial expression database. *Image Vis. Comput.*, 32(10):692–706, 2014. 5

[46] Kaili Zhao, Wen-Sheng Chu, Fernando De la Torre, Jeffrey F Cohn, and Honggang Zhang. Joint patch and multi-label learning for facial action unit detection. In *Proc. IEEE CVPR*, pages 2207–2216, 2015. 2, 6, 7

[47] Kaili Zhao, Wen-Sheng Chu, and Honggang Zhang. Deep region and multi-label learning for facial action unit detection. In *Proc. IEEE CVPR*, pages 3391–3399, 2016. 2, 6, 7

[48] Lin Zhong, Qingshan Liu, Peng Yang, Junzhou Huang, and Dimitris N Metaxas. Learning multiscale active facial patches for expression analysis. *IEEE Trans. Cybern.*, (8):1499–1510, 2015. 2, 6, 7