PIAP-DF: Pixel-Interested and Anti Person-Specific Facial Action Unit Detection Net with Discrete Feedback Learning

Yang Tang[1] Wangding Zeng[1] Dafei Zhao[2] Honggang Zhang[1]
[1]Beijing University of Posts and Telecommunications [2]University of Ottawa
[1]{ty, zengwd, zhhg}@bupt.edu.cn [2]dafei.zhao@uottawa.ca

Abstract

Facial Action Units (AUs) are of great significance in communication. Automatic AU detection can improve the understanding of psychological conditions and emotional status. Recently, several deep learning methods have been proposed to detect AUs automatically. However, several challenges, such as poor extraction of fine-grained and robust local AUs information, model overfitting on person-specific features, as well as the limitation of datasets with wrong labels, remain to be addressed. In this paper, we propose a joint strategy called PIAP-DF to solve these problems, which involves 1) a multi-stage Pixel-Interested learning method with pixel-level attention for each AU; 2) an Anti Person-Specific method aiming to eliminate features associated with any individual as much as possible; 3) a semi-supervised learning method with Discrete Feedback, designed to effectively utilize unlabeled data and mitigate the negative impacts of wrong labels. Experimental results on the two popular AU detection datasets BP4D and DISFA prove that PIAP-DF can be the new state-of-the-art method. Compared with the current best method, PIAP-DF improves the average F1 score by 3.2% on BP4D and by 0.5% on DISFA. All modules of PIAP-DF can be easily removed after training to obtain a lightweight model for practical application.

1. Introduction

Facial expression, a natural way of human communication in people's daily lives, is also an intuitive reflection of human emotions, mental states, and consciousness when analyzing emotion recognition tasks. There are some popular facial expression topics categorized as microexpressions. Microexpressions are reflected by rapid and unconscious spontaneous facial movements, and studies have shown that microexpressions cannot be concealed [8]. These characteristics make the detection of microexpressions necessary in some specific situations, such as the diagnosis of depressed patients [9] and conversations of criminals. Moreover, microexpression detection also has a potential value in many other emotion recognition tasks [15,32,35,43]. In previous studies, the Facial Action Coding System (FACS) [13] method is often used to encode microexpressions. In FACS, each expression is considered as a combination of multiple action units (AU). By detecting the AU, FACS can effectively eliminate the ambiguity problem in microexpression annotation. Therefore, a reliable AU detection system is of great importance for the analysis of facial microexpressions.

In FACS, different AUs are associated with specific facial muscles, which in turn correspond to the features of different regions of the face. Sometimes one AU may also correspond to more than one region. Therefore, local information is essential for AU detection. Traditional [3,7,10,19,26,42] approaches use manual methods to represent different local regions . In recent years, deep learning methods for facial expression detection are gaining popularity and some results have been achieved. Early work used simple CNN for learning. Later, deeper neural networks were employed to improve performance. Due to the im-

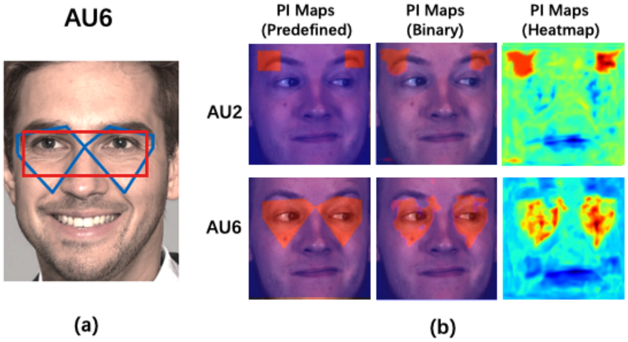

Figure 1. a) Comparison of patch and PI method on AU6, marked as the red border and the blue border, respectively. b) PI Maps on AU2 and AU6. The predefined PI Maps are generated with landmark information. After the second stage of PI, we have refined PI Maps, shown as binary and the heatmap views.

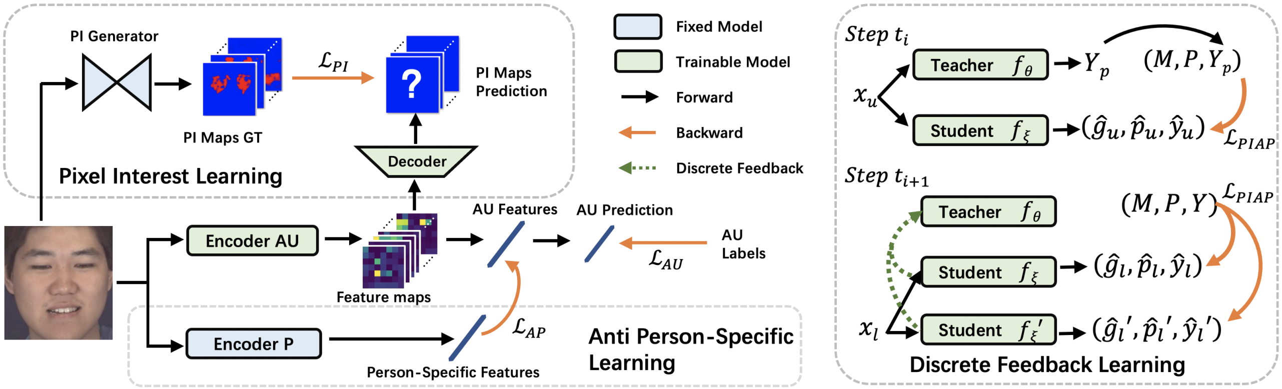

Figure 2. Overview of PIAP and Semi-Supervised Learning of Discrete Feedback.

portance of local features for facial AU detection, previous works typically use facial landmarks to localize these regions or divide the face into patches. In practice, however, AU annotators are sometimes unable to give the exact region of the AU. In other words, the artificially defined AU region correlations are actually not robust prior knowledge. Besides, the giant Patch does not fit the AU-related region well. As shown in Figure 1, these regions are not always rectangles, such as AU6, nor fixed due to uncertainties in the head pose and other factors. In addition, as mentioned earlier, some AUs are simultaneously associated with multiple and fine-grained regions. Therefore, the idea of using a simple landmark-based patch is not very effective. In addition, AU detection should be independent of any specific individual. Due to the limited number of participants in current AU datasets, trained models can be poorly generalized. Therefore, it is necessary to remove person-specific effects on the model.

Recently, self-supervised and semi-supervised learning has made huge leaps. Contrastive learning usually makes the model capable of outputting approximate encodings for different views of the same sample and distinguishing between encodings of different samples, while pseudo-based methods expect models to output low-entropy predictions on the samples never been seen before. It has been proven that self-supervised and semi-supervised learning be effective in improving the generalization ability of the model.

In this paper, we propose PIAP-DF, a set of comprehensive policies for AU detection networks. PIAP integrates two learning strategies, Pixel-Interested (PI) learning and Anti Person-Specific (AP) learning. PI is devoted to providing pixel-level attention for each AU, whereas AP tries to remove person-specific features. Besides PIAP, we also propose Discrete Feedback (DF) technique based on semi-supervised learning, which aims to reduce the effect of mislabeling and improve the robustness. We use EfficientNet-B1 [39] as our AU encoder. In our architecture, PI, AP, and DF can be removed after training to obtain a lightweight network for real-world scenarios.

The main contributions of this paper: 1) We propose a Pixel-Interested learning method to improve the performance of AU detection. PI ensures that the irregular local information and pixel-level correlation of AUs can be retained in the deep layer during forward propagation, thus providing effective supervision of AU detection. 2) We propose an Anti Person-Specific learning method. We eliminate person-specific features from the hidden layer of the network with the help of the same encoder trained on the facial recognition dataset. AP allows the model to focus more on the features of the AU itself on a limited dataset of participants, improving the generality of the network. 3) Based on the characteristics of the dataset and task, we propose a semi-supervised learning strategy with discrete feedback. By utilizing an appropriate amount of additional data and randomly inactivated labels, DF could reduce the impact of mislabeling on training and improve network robustness.

2. Related Works

Facial Action Unit Detection has been studied for decades, and many representative methods have been proposed. In early studies, features like Histogram of Oriented Gradient (HOG) are extracted from the image, and the AU classifier is trained on the extracted features [11,40,49]. For example, Baltrusaitis *et al.* [1] proposed an AU detection method based on the Support Vector Machine (SVM) classifier, which is trained on the HOG features from images after principal component analysis. Since AU is defined as being associated with the movement of facial muscles, many methods also detect the occurrence of AU based on location. Zhao *et al.* proposed JPML [49] to use patches for feature extraction from local regions and a multi-label classifier for AU detection.

Traditional methods are overly dependent on feature extraction and have limitations for complex facial representations. Recently, convolutional neural networks (CNN) have performed well and achieved state-of-the-art on many computer vision tasks, such as object detection [33, 34], classification [18, 21], face recognition [6, 36], and landmark detection [46], etc. This technique has also been introduced to facial AU detection. DRML proposed by Zhao *et al*. [50] adopts region learning to construct highlight regions, achieving the purpose of focusing on local features.

However, AUs consist of the movements of all facial muscles together, and the over-emphasis on local features may cause the global associations to be lost. Due to this limitation, many global feature-based methods have been proposed. Li *et al*. [22] extracts inputs from different parts of the face and merges them into a global feature for AU detection. Their later work [23] further proposes cropping layers to get better local patches. Corneanu *et al*. proposes DSIN [4]. DSIN first extracts features by CNN to initialize predictions for each independent AU, then improves the performance by considering the correlation between AUs.

Because of both the diversity and strong correlation of face tasks, Shao *et al*. [37, 38] proposed JAA and JÂA with joint face alignment for AU detection and multi-task learning. Human faces are person-specific. However, for AU detection, such person-specific features should be eliminated, since the detector is supposed to have similar performance on different faces. Niu *et al*. [29] argue that the facial landmark contains person-specific features, which can be eliminated by making the AU feature orthogonal to the normalized landmark vector. However, our experiments demonstrate that the opposite is true. The features extracted by the landmark detection task contain almost no person-specific features. Consequently, we propose a more robust way to eliminate person-specific features, by making the AU features orthogonal to the features extracted from the same encoder trained on the facial recognition task. Experiments show that this is an effective solution.

Recently, semi-supervised and self-supervised methods [2, 14, 16, 17] have attracted a wide range of attention. Pham *et al*. [31] achieved a new state of the art on ImageNet [5] using semi-supervised methods. For AU detection tasks, Li *et al*. proposed TCAE [24], a twin network method to swap AU features of the source and the target. TCAE is trained on unlabeled data and could achieve performance comparable to supervised learning. Therefore, we further explore the semi-supervised learning approach based on our PIAP. After applying all our strategies simultaneously, we reached a new state of the art.

3. Proposed Method

In this section, we describe our method in detail, including the stages of Pixel-Interested learning, Anti Person-Specific learning, and semi-supervised learning with discrete feedback.

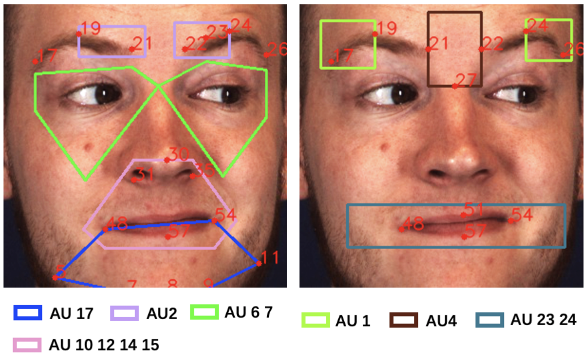

Figure 3. Predefined PI Maps. Red points are facial landmarks indexed from 0 to 67.

3.1. Pixel-Interested Learning

Facial AU detection is a task strongly related to feature location, and some previous works have attempted to associate feature locations to the model for better AU predictions. Most of these methods [22, 23, 49, 50] manually specify a fixed region for each AU and provide input to each AU classifier using a patch-based method. However, these methods have the following problems: 1) the patch is always a rectangle, which leads to imprecise feature extraction. Instead of being simple rectangles, the region of the AU is usually irregular and discontinuous. As shown in Figure 1(a), AU6 involves the muscles of orbicularis oculi and risorius that corresponding to two irregular facial regions. 2) Defined regions are fixed (not trainable), which makes it hard for them to get rid of the mislabeling issue. In practice, AU annotators can not give exact regions for AUs, making such predefined AU regions not capable of being robust prior knowledge.

To solve these problems, we propose a multi-stage Pixel-Interested (PI) learning, demonstrated in Figure 4. PI involves the following three stages.

In the first stage, based on our understanding of AUs, we manually define the regions for each AU on the basis of the 68-point facial landmarks, naming predefined Pixel Interest Maps (PI Maps). PI Maps act as a binarized mask to block the features outside the region. Figure 3 gives out the definition of predefined PI Maps for each AU based on landmarks. To generate predefined PI Maps without landmark information, we trained a modified UNet on the AFEW-VA dataset [20]. This modified UNet, which we call DW-UNet, replaces the original convolution layers with the depthwise separable convolution and employs the binary cross-entropy loss (BCELoss) as the loss function $\mathcal{L}_{PI}$ (1). Here, H and W represent the height and width of PI-Map; Y denotes the ground truth, and $\hat{Y}$ denotes the predicted value, the same below. DW-UNet reduces the number of parameters and calculations by 90%. The trained DW-UNet will act as PI Generator 0 in the second stage, where we use it to gener-

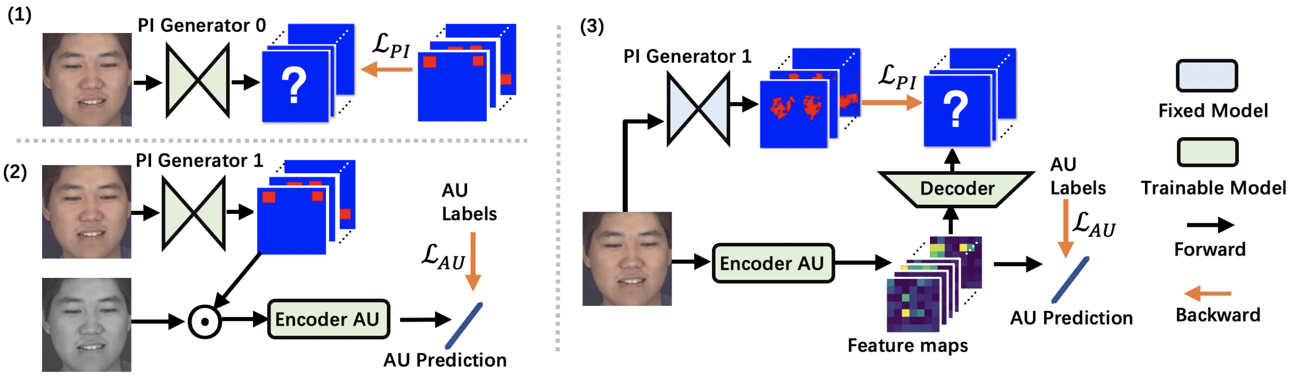

Figure 4. Overview of the 3 stages of Pixel-Interested learning.

ate 12 predefined PI Maps (12 × 256 × 256) from the raw images.

$$\mathcal{L}_{PI} = -\frac{1}{HW}\sum_{H}^{h=1}\sum_{W}^{w=1} Y_{hw}\cdot \log \hat{Y}_{hw} + (1-Y_{hw})\cdot \log(1-\hat{Y}_{hw}) \quad (1)$$

In the second stage, for each image input, we first put the RGB format sample (3 × 256 × 256) into PI Generator 0 to obtain 12 predefined PI Maps. Then we perform Hadamard product on the predefined PI Maps (before One-Hot encoding) and the grayscale format sample to generate 12-channel features, which is supposed to have key regions highlighted. This feature map will be used as the input of the encoder. In this stage, the encoder we used is an EfficientNet-B1 pretrained on ImageNet, with the number of input channels changed to 12 to fit the features. It performs as the actual AU classifier to outputs AU predictions in the end. We train the whole model supervised by $\mathcal{L}_{AU}$ (2) on the AU predictions, where C represents the total number of AU categories. In this process, as the encoder converges, the parameters of the PI generator are also updated, thus facilitating the generator to better find the pixels of interest. As a result, a more robust association between AU and pixel-level region information is established, and we can get a new PI Generator 1 after training. The PI Maps generated by PI Generator 1 are pixel-level, better than the predefined PI Maps (Figure 1). Note the PI Maps are adjusted after training. As a quick reference, the model performance (F1-score, %) at this stage reaches 60.5. We'll exhibit more details in the ablation study.

$$\mathcal{L}_{AU} = -\frac{1}{C}\sum_{C}^{c=1} Y_c \cdot \log \hat{Y}_c + (1-Y_c)\cdot \log(1-\hat{Y}_c) \quad (2)$$

In the third stage, we still use an EfficientNet-B1 model pretrained on ImageNet as the AU encoder. What's different in this stage is that we add a decoder branch before the last average pooling layer, while fixing the parameters of the PI Generator. The decoder consists of DWConv layers and upsampling layers. It accepts the feature map of 1280 × 7 × 7 and is trained to output PI Map predictions as close as the ones (ground truth, 12 × 256 × 256) generated from PI Generator 1. The loss is measured by $\mathcal{L}_{PI}$. In this way, the whole model is supervised by both $\mathcal{L}_{AU}$ and $\mathcal{L}_{PI}$, and we add a hyper-parameter α to adjust their weights, as shown in (3). So what are the benefits of doing this? The features in front layers may be discarded or diffused in the deep layers, making it contain no or sparse location-related information. However, AU detection is a strong location-related task. If we can extract a similar PI Map from the feature map by a decoder, it must contain some degree of location information. As for the location information we used as ground truth, it reflects what the PI Generator learned in stage 2 and could be regarded as robust prior knowledge. This information, in turn, provides additional supervision for AU detection. The model reaches F1-score performance of 63.9% in this stage.

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{AU} + (1-\alpha)\cdot \mathcal{L}_{PI} \quad (3)$$

In summary, we first use predefined PI Maps to train a PI Generator. Then the supervision of the AUs makes the PI Generator learn finer pixel interest information. At last, the new PI Generator is used to sustain pixel-interest information in the deep layers of the network, providing additional and effective supervision for AU detection. In later sections, we will see how Pixel-Interested learning can be used jointly with other strategies to further improve performance.

3.2. Anti Person-Specific Learning

Due to expensive annotation costs and privacy limitations, the AU detection datasets always contain a large number of video frames but very few participants. For example, BP4D dataset has more than 140,000 labeled frames but only 41 participants. Compared with the complex distribution of AU features, the model is more likely to learn person-specific features of different participants and predict AUs based on a specific person. This explains why most models can achieve high F1cores on the training set but perform poorly on the validation and the test set.

To address this problem, LP-Net [29] takes normalized facial landmarks as person-specific features and tries to eliminate their impact by making global AU features orthogonal to it. However, the facial landmark is actually patterned and structured, rather than person-specific information. In other words, a facial landmark detection model works because it knows that Rachel, Ross, and Joey all have eyes and noses, not because it knows Rachel is Rachel. So using facial landmarks as a regularization condition will instead cause the network to remove the structured, patterned, and non-person-specific features needed for AU detection, making the AU detection less effective. We give a toy case here to prove that the face landmark detection task is also not person-specific. We first trained a Resnet50 [18] model on the AFEW-VA dataset [20] and get Mean Square Error (MSE) of 0.6%. Then we use the 1000-dimension vector before the last fully-connected layer as the feature vector to validate on the LFW dataset [27] in the general way [6]. The result shows a low accuracy of 58%, which confirms our idea. We also try to apply the strategy in our experiments, and unsurprisingly the landmark regularization decreases the baseline performance. Due to space limitations, we only describe it briefly here and this toy case can be easily reproduced.

Our solution is Anti Person-Specific learning (AP), as shown in Figure 2. We add an Encoder P with the same model structure as Encoder AU, so that the two encoders have the same parameter space. First, we train the Encoder P for face recognition task on the CASIA-WebFace dataset [45], and after training it achieves 96.2% of accuracy on another dataset LFW [27]. After that, we could use the vector v_P output by Encoder P as person-specific features. To make the AU features independent from the person-specific features, we use $\mathcal{L}_{AP}$ defined in (4) to encourage v_{AU} and v_P to be orthogonal, thus making f_{AU} and f_P orthogonal functions. Here v_{AU} refers to the AU feature vector output by Encoder AU. Ultimately, supervised by $\mathcal{L}$ defined in (5), the features extracted by encoder AU could filter out personal features as much as possible. Here, we use β to adjust the weight between $\mathcal{L}_{AU}$ and $\mathcal{L}_{AP}$. The average F1-score after applying AP to the baseline reaches 61.9 (%).

$$\mathcal{L}_{AP} = -\frac{1}{N}\sum_{N}^{n=1}\log(1 - \frac{|\langle v_P, v_{AU}\rangle|}{\|v_P\|_2 \cdot \|v_{AU}\|_2}) \tag{4}$$

$$\mathcal{L} = \beta \cdot \mathcal{L}_{AU} + (1-\beta) \cdot \mathcal{L}_{AP} \tag{5}$$

3.3. Discrete Feedback Semi-Supervised Learning

Due to the characteristics of the task domain, the existing AU datasets are not yet perfect. First, the BP4D dataset contains a number of incorrect labels. Generally, the noise from incorrect labels in the dataset can obviously affect the performance of the model. Second, both datasets face the problem of insufficient participants, as mentioned in Section 3.2. This problem is particularly evident for DISFA. In previous works, DISFA dataset is usually not used for the full training due to its limited number of participants. As for BP4D, it has more than 200,000 unlabeled frames that also limits the effective utilization.

To mitigate the impact of the above two problems for the model, we want to 1) make full use of samples of DISFA and unlabeled samples of BP4D to improve model's generalizability by semi-supervised learning; 2) randomly pick up labeled data from the BP4D dataset as unlabeled samples to iteratively encourage the semi-converged model to converge in the right direction against incorrect labels. The latter is actually another kind of regularization strategy. Inspired by feedback strategy of MPL [31], we propose Discrete Feedback strategy to improve the performance of semi-supervised learning.

As shown in Figure 2, well pretrained PI and AP jointly as f. Then f is copied to f_θ and f_ξ, where f_θ is the Teacher and f_ξ is the Student. Let their parameters be θ and ξ. We use x_u and x_l to refer to unlabeled and labeled samples, respectively. Let Y be the labels, and Y_p be the pseudo labels generated by Teacher. We use $G_l = (M, P, Y)$ and $G_p = (M, P, Y_p)$ to refer to these ground truth pairs, where M is the ground truth PI Maps and P is person-specific features. We denote $(\hat{m}, \hat{p}, \hat{y})$ as the prediction pairs and let $\hat{Y}_{\theta,l}$ be the value of $f_\theta(x_l)$. $\mathcal{L}_{PIAP}$ is defined as (6), where α, β, and γ are used to adjust the weights of each components.

$$\begin{cases} \mathcal{L}_{PIAP} = \alpha \cdot \mathcal{L}_{AU} + \beta \cdot \mathcal{L}_{PI} + \gamma \cdot \mathcal{L}_{AP} \\ \alpha + \beta + \gamma = 1 \end{cases} \tag{6}$$

Every step t_i, unlabeled sample x_u is used to update θ. x_u contains unlabeled samples and few labeled samples of BP4D and samples of DISFA without labels. $\hat{Y}_{\theta,u}$ refers to predictions of f_θ based on unlabeled data. g refers to gradient.

$$\begin{aligned} g_{\theta,u} &= \nabla_\theta \mathcal{L}_{PIAP}(\hat{Y}_{\theta,u}, G_p) \\ g_{\xi,u} &= \nabla_\xi \mathcal{L}_{PIAP}(\hat{Y}_{\xi,u}, G_p) \end{aligned} \tag{7}$$

$$\xi' = \xi - \eta g_{\xi,u} \tag{8}$$

In each step t_{i+1}, labeled sample x_l is used to compute feedback fb. Then θ is updated by σ, $fb_{Discrete}$, $g_{\theta,u}$ and $g_{\theta,l}$. In the training process, we found the feedback fb may be too large or too small due to the accumulation of SGD's momentum and other factors in the model such as Dropout. Besides, some feedback is likely to be wrong because of the wrong labels. Consequently, we proposed discrete feedback learning. We make the feedback as discrete format and randomly disable it by chance σ (10) to weaken the negative impact. When the Student gives positive feedback, $fb_{Discrete}$ is set to 1, or -0.1 for negative feedback, and 0 for σ of positive. Where σ is a random factor, with a 5% possibility of being 0 otherwise 1. This strategy performs better than MPL in our task.

$$\begin{aligned} g_{\theta,l} &= \nabla_{\theta}\mathcal{L}_{PIAP}(\hat{Y}_{\theta,l}, G_l) \\ g_{\xi',l} &= \nabla_{\xi'}\mathcal{L}_{PIAP}(\hat{Y}_{\xi',l}, G_l) \end{aligned} \tag{9}$$

$$\begin{aligned} fb &= {g_{\xi',l}}^{\top} \cdot g_{\xi,u} \\ fb_{Discrete} &= \sigma \cdot h(fb) \\ h(z) &= \begin{cases} 1, z > 0 \\ -0.1, z \leqslant 0 \end{cases} \end{aligned} \tag{10}$$

$$\theta' = \theta - \eta \cdot (fb_{Discrete} \cdot g_{\theta,u} + g_{\theta,l}) \tag{11}$$

$$\begin{aligned} \theta^{+} &= \arg\min_{\theta} \mathbb{E}_{x_l,G_l}[\mathcal{L}_{PIAP}(\hat{Y}_{\xi^{+},u}, G_l)] \\ \xi^{+} &= \arg\min_{\xi} \mathbb{E}_{x_u}[\mathcal{L}_{PIAP}(\hat{Y}_{\xi,u}, G_p)] \end{aligned} \tag{12}$$

For each step, according to [12, 25], ∇_{θ} can be deducted as

$$\begin{aligned} \nabla_{\theta} &= \frac{\partial \mathbb{E}_{x_u,G_p}\mathcal{L}_{PIAP}(\hat{Y}_{\xi',l}, G_l)}{\partial\theta} \\ &= \frac{\partial \mathcal{L}_{PIAP}(\hat{Y}_{\xi',l}, G_l)}{\partial\xi'} \cdot \frac{\partial \mathbb{E}_{x_u,G_p}[\xi']}{\partial\theta} \\ &= g_{\xi',l} \cdot \frac{\partial \mathbb{E}_{x_u,G_p}[\xi - \eta \cdot \nabla_{\xi}\mathcal{L}_{PIAP}(\hat{Y}_{\xi,u}, G_p))]}{\partial\theta} \end{aligned} \tag{13}$$

and can be further calculated as the following [44]:

$$\begin{aligned} \nabla_{\theta} &= \eta \cdot g_{\xi',l} \cdot \frac{\partial \mathcal{L}_{PIAP}(\hat{Y}_{\xi,u}, G_p)}{\partial\xi} \cdot -\frac{\partial \log(P(G_p))}{\partial\theta} \\ &= \underbrace{\eta \cdot {g_{\xi',l}}^{\top} \cdot g_{\xi,u}}_{feedback:fb} \cdot g_{\theta,u} \end{aligned} \tag{14}$$

We can find feedback fb in (14). The detailed derivation can be found in Appendix. This process achieves the effect that Teacher may make mistakes, so the feedback given by Student's performance on the labeled data can correct its mistakes. Note that the feedback is randomly disabled to adapt to the presence of incorrect labels.

3.4. Summary of PIAP-DF

To combine all our methods, we first perform the first and second stages of Section 3.1 to obtain the trained PI Generator 1 as the PI Generator in Figure 2. Then we have Encoder P trained as 3.2, and we take $\mathcal{L}_{PIAP}$ to calculate the loss. After that, we make 2 copies of PIAP with PI Generator and Encoder P fixed and Encoder AU initialized, as Teacher and Student. The Teacher and Student then get trained on labeled and unlabeled data as 3.3. This marks the completion of PIAP-DF. The AU Encoder of Student is the final model we need for AU detection.

4. Experiment

In this section, we show the experimental evaluation of PIAP on two widely used AU detection datasets and give the results of ablation experiments on BP4D to investigate the effectiveness of PI, PA and DF. We also give the results of PI Maps for some AUs generated by Pixel-Interested learning.

4.1. Experiment Setting

4.1.1 Dataset

AU datasets are much more limited than other image task datasets due to their stringent requirements and the limitations of the task itself. In this paper, we use two widely used AU detection datasets, BP4D and DISFA.

BP4D [48] contains 23 female and 18 male participants. 8 different tasks are tested on the 41 participants, and their spontaneous expressions are recorded in several videos. In the recorded 328 videos, 12 AUs are coded by 0 or 1 without intensity information. There are 140,000 labeled frames and 240,000 unlabeled frames in the 2D video we used in BP4D.

DISFA [28] involves 27 participants, 12 females and 15 males. Each participant is asked to watch a video, and their facial features are recorded during the process. DISFA contains more than 100,000 video frames with 12 AU labels ranging from [0, 5], in which 8 AU labels are used for experimental comparisons. We use 2 as a threshold to distinguish between positive and negative samples.

4.1.2 Training

We train our model on the two datasets with slight differences. On the BP4D dataset, we train the model with 3-fold cross-validation to verify the validity and universality of the method. The division of the dataset is based on the participants' IDs. On the DISFA dataset, since we have no additional unlabeled data available, we do not perform semi-supervised learning on it. As the well-trained PIAP-DF is based on the DISFA dataset, we alternatively fine-tune the PIAP model trained on BP4D to evaluate the performance. We extract the 1280-dimension vector before the original fully connected layer and reconnect it to an 8-dimension

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 [49]	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
SplitBrain [47]	39.0	32.0	39.7	72.9	70.6	78.2	83.7	57.8	37.3	53.6	32.3	45.1	53.5
TCAE [24]	43.1	32.2	44.4	75.1	70.5	80.8	85.5	61.8	43.7	58.5	37.2	48.7	56.1
DRML [50]	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
ROI [22]	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
DSIN [4]	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
EAC-Net [23]	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
LP-Net [29]	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
JAA-Net [37]	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
JÂA-Net [38]	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
PIAP	**54.2**	47.1	54.0	**79.0**	**78.2**	**[86.3]**	**89.5**	**[66.1]**	**49.7**	**63.2**	**49.9**	**52.0**	**64.1**
PIAP-DF	**[55.0]**	**[50.3]**	51.2	**[80.0]**	**[79.7]**	84.7	**[90.1]**	65.6	**[51.4]**	**[63.8]**	**[50.5]**	50.9	**[64.4]**

Table 1. Comparison of F1 score (in %) on BP4D Dataset. Bracketed and bold numbers indicate the best performance; bold numbers indicate the second best.

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
SplitBrain [47]	13.1	10.6	35.7	40.2	30.2	57.5	77.4	40.3	38.1
TCAE [24]	15.1	15.2	50.5	48.7	23.3	72.1	82.1	52.9	45.0
DRML [50]	17.3	17.7	37.4	29.0	10.7	37.7	38.5	20.1	26.7
ROI [22]	41.5	26.4	66.4	**[50.7]**	8.5	**[89.3]**	88.9	15.6	48.5
JAA-Net [37]	43.7	46.2	56.0	41.4	44.7	69.6	88.3	58.4	56.0
DSIN [4]	42.4	39.0	68.4	28.6	46.8	70.8	90.4	42.2	53.6
LP-Net [29]	29.9	24.7	**[72.7]**	46.8	**49.6**	72.9	**93.8**	**65.0**	56.9
JÂA-Net [38]	**[62.4]**	**[60.7]**	67.1	41.1	45.1	73.5	90.9	**[67.4]**	**63.5**
PIAP	**50.2**	**51.8**	**71.9**	**50.6**	**[54.5]**	**79.7**	**[94.1]**	57.2	**[63.8]**

Table 2. Comparison of F1 score (in %) on DISFA Dataset. Bracketed and bold numbers indicate the best performance; bold numbers indicate the second best.

output vector to obtain the new AU prediction. In this work, all the implementations are based on PyTorch [30].

4.1.3 Metric

AU detection is a multi-label binary classification task, where F1-score can be a good metric. In previous works, F1-score is also a common evaluation criterion [10, 41]. We calculate F1-score for 12 AUs in BP4D and 8 AUs in DISFA. F1-score can be directly compared as a performance indicator for different algorithms on each AU.

4.2. Comparison

For the sake of rigorous and reasonable experimental results, we compare PIAP with the current image-based AU detection methods using 3-fold cross-validation, including the traditional methods LSVM [11], JPML [49], and the supervised methods DRML [50], ROI [22], DSIN [4], EAC-net [23], LP-Net [29], JAANet [37], JÂANet [38], and two semi-supervised methods TCAE [24] and SplitBrain [47].

Table 1 shows the performance comparison of PIAP with other AU detection methods on BP4D. Overall, PIAP shows excellent performance on this widely used AU detection dataset. Compared to the existing best method, JÂAnet, PIAP with Discrete Feedback learning (PIAP-DF) achieves an average performance improvement of 3.2% for F1-score.

Table 2 shows the experimental results of PIAP on the DISFA dataset. PIAP improves the F1 score by 0.5% on average over the best method. For DIFSA, we do not train the model from scratch, but fine-tune the model trained on BP4D. For this reason, our strategy cannot be directly applied to DISFA, resulting in weaker improvements for each AU than that on BP4D, but it still outperforms the existing automatic AU detection methods. This also indicates an improved generalization ability of our model. On BP4D which directly benefited from it, the model performance on most individual AUs could achieve the best.

Method	**AU1**	**AU2**	**AU4**	**AU6**	**AU7**	**AU10**	**AU12**	**AU14**	**AU15**	**AU17**	**AU23**	**AU24**	**Avg.**
Baseline	44.3	42.8	51.9	71.5	68.6	80.7	80.8	57.9	49.4	54.9	43.5	45.8	57.7
PI-2	44.7	21.8	48.3	75.2	78.7	81.9	82.7	65.4	67.2	61.9	47.7	50.1	60.5
PI-3	**54.6**	44.1	**[56.2]**	77.8	77.8	**85.5**	88.9	**65.9**	**50.7**	**63.5**	**50.1**	**51.0**	63.9
AP	54.1	**49.2**	48.8	75.7	**79.6**	83.7	89.3	63.2	45.3	61.8	48.7	42.9	61.9
DF	46.8	43.3	51.2	77.8	71.4	82.1	87.0	61.2	45.1	59.9	41.9	44.8	59.4
PIAP	54.2	47.1	**54.0**	**79.0**	78.2	**[86.3]**	**89.5**	**[66.1]**	49.7	63.2	49.9	**[52]**	**64.1**
PIAP-DF	**[55.0]**	**[50.3]**	51.2	**[80]**	**[79.7]**	84.7	**[90.1]**	65.6	**[51.4]**	**[63.8]**	**[50.5]**	50.9	**[64.4]**

Table 3. F1-score (in %) of ablation experiments on the BP4D dataset. Bracketed bold number indicates the best performance, and bold numbers for second best. **PI-2**: Baseline with stage-2 of PI; **PI-3**: Baseline with stage-3 of PI; **AP**: Baseline with AP; **DF**: Baseline with DF; **PIAP**: Baseline with stage-3 of PI, and AP; **PIAP-DF**: Baseline with PIAP and DF.

From the results, benefiting from the combined use of the three strategies, PIAP-DF shows excellent performance. In terms of actual deploying, the training module PI, AP, and DF can all be easily removed to get a lightweight inference model. Note this lightweight model does not need any additional information such as facial landmarks. These advantages make PIAP-DF significantly better than other methods.

4.3. Ablation Study

In this section, we demonstrate the ablation study on PIAP-DF to investigate the effectiveness of Pixel-Interest learning (PI), Anti Person-Specific learning (AP), and Discrete Feedback Semi-Supervised Learning (DF). Table 3 shows the F1-score by individual ablation experiments on BP4D. All of the results are based on 3-fold cross-validation experiments.

We use EfficientNet-B1 as the baseline. After we apply the first 2 stages of PI (PI-2)to the baseline (the first stage of PI does not output AU predictions), the average F1-score reaches 60.5%, an improvement of 4.9% points from baseline. Stage-3 of PI (PI-3 or PI) further improves on PI-2 by 5.6% to 63.9%. If we apply both PI and AP on the baseline, the PIAP can achieve an 11.1% improvement compared to baseline, reaching an average F1-score of 64.1%. For DF, it can be used alone on the baseline to achieve a score of 59.3% and gain a 2.8% increase, or combine it with PIAP together, forming the complete PIAP-DF, to get the best F1-score of 64.4%. This is also the current state-of-the-art method as we know.

The ablation study proves the individual strategies we designed further improve the performance.

4.4. Results of Pixel-Interested Learning

In this section, we show the PI Maps of AU2 and AU6 generated by the PI Generator in PI-2, and compare them with the predefined PI Maps in Figure 1. We use a translucent mask to refer to the predefined PI Map used in the traditional methods and the binary PI Map used in PI-2 method, where the red pixels are 1 and the blue pixels are 0. We also show the heatmap-format PI Maps in PI-2, where the value of red pixels is larger than the blue and green ones. By comparison, we can see that PI Generator generates more refined PI Maps, unlike patches or predefined PI Maps that must have straight edges and regular shapes. These more refined PI Maps also proved to be more effective in our ablation study. By presenting these PI Maps, we hope that the pixel-level and multi-region PI Maps can give some useful inspiration to AU annotators and researchers.

5. Conclusion

Automatic face AU detection is a challenging task. In this paper, we propose an integrated strategy approach incorporating: a strategy PI for pixel-level interest learning, a strategy AP for person-specific information removal, and a semi-supervised learning method with discrete feedback. These methods can be used jointly to train models for AU detection tasks on a limited dataset with incorrect labels. By evaluating on two generic AU datasets, PIAP-DF makes the final model outperform all existing models. At last, PIAP-DF, as a flexible training strategy, any model trained with it can be easily exported as a lightweight AU encoder for inference purposes. These encoders can better fit the production environment, such as mobile or IoT devices.

6. Acknowledgement

This work was supported in part by the National Natural Science Foundation of China (NSFC) under joint grants 62076034 and 61806184.

References

[1] Tadas Baltrušaitis, Marwa Mahmoud, and Peter Robinson. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 6, pages 1–6. IEEE, 2015.

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. 2020.

[3] Wen-Sheng Chu, Fernando De la Torre, and Jeffery F Cohn. Selective transfer machine for personalized facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3515–3522, 2013.

[4] Ciprian Corneanu, Meysam Madadi, and Sergio Escalera. Deep structure inference network for facial action unit recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 298–313, 2018.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[6] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[7] Xiaoyu Ding, Wen-Sheng Chu, Fernando De la Torre, Jeffery F Cohn, and Qiao Wang. Facial action unit event detection by cascade of tasks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2400–2407, 2013.

[8] Paul Ekman. Darwin, deception, and facial expression. *Annals of the New York Academy of Sciences*, 1000(1):205–221, 2003.

[9] Paul Ekman and Wallace V Friesen. Nonverbal leakage and clues to deception. *Psychiatry*, 32(1):88–106, 1969.

[10] Stefanos Eleftheriadis, Ognjen Rudovic, and Maja Pantic. Multi-conditional latent variable model for joint facial action unit detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3792–3800, 2015.

[11] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874, 2008.

[12] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017.

[13] E Friesen and Paul Ekman. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto*, 3(2):5, 1978.

[14] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

[15] John Mordechai Gottman and Robert Wayne Levenson. A two-factor model for predicting when a couple will divorce: Exploratory analyses using 14-year longitudinal data. *Family process*, 41(1):83–96, 2002.

[16] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[19] Sander Koelstra, Maja Pantic, and Ioannis Patras. A dynamic texture-based approach to recognition of facial actions and their temporal models. *IEEE transactions on pattern analysis and machine intelligence*, 32(11):1940–1954, 2010.

[20] Jean Kossaifi, Georgios Tzimiropoulos, Sinisa Todorovic, and Maja Pantic. Afew-va database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, 65:23–36, 2017.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[22] Wei Li, Farnaz Abtahi, and Zhigang Zhu. Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1841–1850, 2017.

[23] Wei Li, Farnaz Abtahi, Zhigang Zhu, and Lijun Yin. Eacnet: Deep nets with enhancing and cropping for facial action unit detection. *IEEE transactions on pattern analysis and machine intelligence*, 40(11):2583–2596, 2018.

[24] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Self-supervised representation learning from videos for facial action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10924–10933, 2019.

[25] Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*, 2018.

[26] Mengyi Liu, Shaoxin Li, Shiguang Shan, and Xilin Chen. Au-aware deep networks for facial expression recognition. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6. IEEE, 2013.

[27] Peter M. Roth Martin Koestinger, Paul Wohlhart and Horst Bischof. Annotated Facial Landmarks in the Wild: A Large-scale, Real-world Database for Facial Landmark Localization. In *Proc. First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011.

[28] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013.

[29] Xuesong Niu, Hu Han, Songfan Yang, Yan Huang, and Shiguang 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*, pages 11917–11926, 2019.

[30] 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. 2017.

[31] Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V Le. Meta pseudo labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11557–11568, 2021.

[32] Lorraine Dacre Pool and Pamela Qualter. Improving emotional intelligence and emotional self-efficacy through a teaching intervention for university students. *Learning and Individual Differences*, 22(3):306–312, 2012.

[33] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[34] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015.

[35] Frank Salter, Karl Grammer, and Anja Rikowski. Sex differences in negotiating with powerful males. *Human Nature*, 16(3):306–321, 2005.

[36] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[37] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. Deep adaptive attention for joint facial action unit detection and face alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 705–720, 2018.

[38] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. Jâa-net: Joint facial action unit detection and face alignment via adaptive attention. *International Journal of Computer Vision*, 2020.

[39] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019.

[40] Michel Valstar and Maja Pantic. Fully automatic facial action unit detection and temporal analysis. In *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, pages 149–149. IEEE, 2006.

[41] Michel F Valstar, Timur Almaev, Jeffrey M Girard, Gary McKeown, Marc Mehu, Lijun Yin, Maja Pantic, and Jeffrey F Cohn. Fera 2015-second facial expression recognition and analysis challenge. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 6, pages 1–8. IEEE, 2015.

[42] Ziheng Wang, Yongqiang Li, Shangfei Wang, and Qiang Ji. Capturing global semantic relationships for facial action unit recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3304–3311, 2013.

[43] Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster, and Javier R Movellan. The faces of engagement: Automatic recognition of student engagementfrom facial expressions. *IEEE Transactions on Affective Computing*, 5(1):86–98, 2014.

[44] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

[45] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[46] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[47] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1058–1067, 2017.

[48] 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 and Vision Computing*, 32(10):692–706, 2014.

[49] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2207–2216, 2015.

[50] Kaili Zhao, Wen-Sheng Chu, and Honggang Zhang. Deep region and multi-label learning for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3391–3399, 2016.