IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 20, 2021, accepted February 18, 2021, date of publication March 29, 2021, date of current version April 7, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3069468

Facial Expression Recognition With Confidence Guided Refined Horizontal Pyramid Network

WEN SU[1], HAIFENG ZHANG[2], YUAN SU[3], AND JUN YU[2]

[1]Faculty of Mechanical Engineering and Automation, Zhejiang Sci-Tech University, Hangzhou 310018, China
[2]Department of Automation, University of Science and Technology of China, Hefei 230016, China
[3]School of Optoelectronic Engineering, Xi'an University of Technology, Xi'an 710000, China

Corresponding author: Wen Su (wensu@zstu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 62006209, in part by the Zhejiang Natural Science Fund Project of China under Grant LQ20F020001, in part by the Science Foundation of Zhejiang Sci-Tech University under Grant 18022225-Y, and in part by the Fundamental Research Funds of Zhejiang Sci-Tech University under Grant 2020Q014.

ABSTRACT Facial expression recognition has become one of the most studied applications in computer vision and human-computer interaction. Part-level features constitute one the most appealing breakthrough to offer fine-grained information and have been intensively studied consequently. The main prerequisite is accurate location of facial parts with additional cues, which is turned to be another difficulty. Rather than directly locating parts, this paper proposes a confidence guided Refined Horizontal Pyramid Network (RHPN) to fully exploit various partial information of a given facial image. It can learn discriminative features with multiple facial granularities. Inconsistencies within each granularity are efficiently limited. Specifically, we first design a Horizontal Pyramid Network (HPN) for classification using the uniform partial feature representations at different horizontal pyramid scales. It successfully enhances the discriminative capabilities of various facial parts. Then, a refinement mechanism is added to HPN due to the fact that the uniform partition inevitably incurs outliers in each part and introduces unreasonable similarity between different parts. It re-assigns these outliers to the parts they are closest to, resulting in the refined parts with enhanced within-stripe consistency. Due to the lack of explicit supervisory information, we design an induced training strategy additionally for efficiently training of RHPN. For avoiding the finagle of prediction based on the separate feature vector, we prefer the confidence guided prediction as the final classification result. Experiments verify the effectiveness of our network for increasing the performance on facial expression recognition. More importantly, it surpasses the state-of-the-art by a large margin on the lab-controlled Oulu-CASIA dataset and the real world RAF-DB dataset.

INDEX TERMS Facial expression recognition, local feature learning, content consistency, multiple granularities.

I. INTRODUCTION

Facial expressions play significant roles in our daily communication. People can not only convey their inner thoughts through facial expressions, but also understand each other through facial expressions. Consequently, recognizing facial expressions has become an fundamental research in extensive applications such as human-computer interaction [1], multimedia [2], and security [3], [4].

Although astonishing progress based Convolutional Neural Networks (CNNs) has been made [5]–[7] in recent years, facial expression recognition is still a challenging problem. The expression recognition confronts with large variations of face, including variant poses, illumination variations, occlusions and identity bias, etc. This led researchers to seek approaches which can hold global expression related discriminates and are robust to local variations caused by non-expressions. The intuitive methods are based on discriminative global features extracting from the entire facial image. Recently, some global-based methods focus on learning metrics or extract features from entire facial image [8]–[13]. Their purpose is to capture the most significant appearance cues to distinguish facial expressions. However, due to the complexities of visual cues in facial expression

The associate editor coordinating the review of this manuscript and approving it for publication was Tallha Akram.

images, methods that rely solely on global features tend to be less robust. More importantly, the differences between individuals with the same expression may beat the subtle differences between different expressions leading to a limited accuracy of facial expression recognition with global representations. When extracting features from global facial image directly through CNNs, some insignificant or infrequently detailed information is easily overlooked. When different expressions have extremely similar appearances, such detailed information is particularly important for accurately distinguishing them. In other words, as a task which has small inter-class differences and large intra-class differences, global features based methods are difficult to adapt to the classification of facial expression recognition.

In order to solve this problem, it has been proved to be an effective way to extract local information of facial expressions from the local region of the facial image. Each local region encodes subtle representation of the corresponding facial position. The features from the local region can filter the interference of other related or irrelevant information outside the current region. These local regions can be used to learn more hidden crucial features. Some recent approaches have been focusing on learning partial discriminative feature representations. The part-based methods can be roughly divided into two categories according to how they acquire local features: 1) Locate the local region-of-interest (ROI) and extract the local features of the corresponding area [14]–[16]. They need to leverage datasets beyond facial expression and complex detector to establish a localization model for locating the ROI, including external supervision signals such as facial landmarks, facial pixel-level segmentation and image level labels. Then, these localization models are all transferred to the facial expression dataset. The potential deviations between facial expression dataset and the third-party dataset will inevitably hinder the accurate localization of local region on facial expression images. Moreover, if only focusing on the features of the ROI, it cannot cover all the discriminating information. 2) Without regional positioning, the facial expression image is cropped into partial patches [17], [18]. Due to the diversity of data, the size of local patches is difficult to determine. Too large local patches will make it difficult to extract the detailed information effectively, while too small patches will make it difficult to guarantee the integrity of the local features. Moreover, introducing all local patches without selection is time-consuming and inefficient. These patches may not be optimal for the final classification. Some patches may have no means or even have a negative impact for facial expression recognition, and thus prone to errors introduced by outliers.

For above problems, this paper proposes confidence guided Refined Horizontal Pyramid Network (RHPN) and induced training strategy. The network is constructed by horizontal pyramid network (HPN) and refine mechanism. First, without any additional local supervision, the proposed HPN simultaneously exploits global features and multi-granularity local features through uniformly partitioned facial image parts.

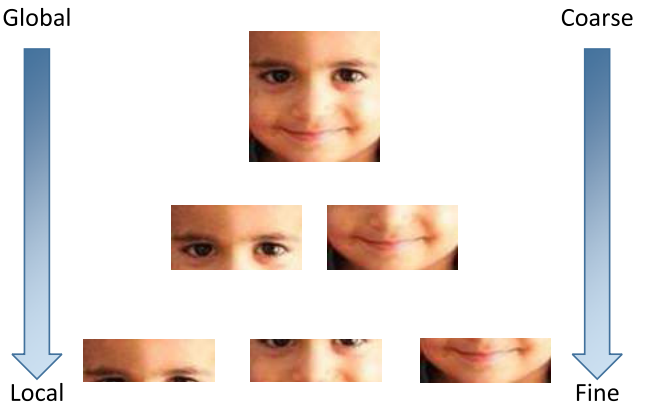


FIGURE 1. Facial partitions from global to local and coarse to fine.

As shown in Figure 1, according to the physiological structure of human face, the facial image can be divided into uniform stripes horizontally. They are segmented according to different granularities. We take the original image containing the entire face in the top row as the coarsest level of granularity. The middle row and the bottom row are the face partitions divided into 2 and 3 from the original image, respectively. The more stripes are divided, the finer the granularity of the region is. Different numbers of uniform feature stripes introduce diversity in content granularity. HPN employs a feature learning strategy that combines global information and local information. It also generates representations with coarse to fine granularities. As shown in Figure 2, to obtain local features with multiple granularities, we uniformly divide the feature of backbone network into different feature stripes instead of explicitly dividing the image into local patches. Second, based on the local features extracted by HPN, there are outliers in the uniformly divided partitioned parts. The features that are divided into a certain region are closer to the features of another region, and the presence of the outliers interferes with the discrimination of the features of each local region. Therefore, we add a refinement mechanism to HPN. A structure called refinement network is added to the second and third branches of HPN to emphasize the consistency within each partitioned part. Through the designed induced training procedure, the refined network is successfully used to adjust the outliers in the partitioned part. The confidence-guided prediction can avoid the finagle of prediction based on the separate feature vector so that it makes the whole model robust to local region disturbances. Specifically, we make the following contributions:

- We propose Horizontal Pyramid Network (HPN) which uniform divides the deep feature maps into different numbers of spatial stripes with various pyramid scales. Each spatial stripe feature is then used for independent classification. Various numbers of partitioned stripes introduce different feature granularities. If we define global feature are extracted from the original image that contains whole partition, as the number of partitions

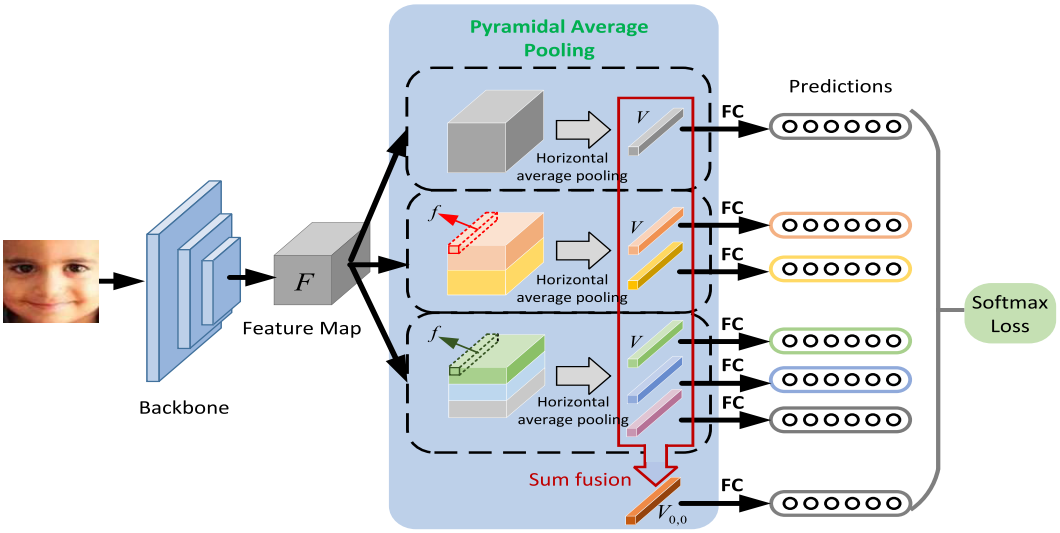


FIGURE 2. Overview of the proposed Horizontal Pyramid Network (HPN). A backbone is employed to extract feature maps. Then, the pyramidal average pooling is employed to producing feature vector of each stripe. Each feature vector is fed to the classifier for expression recognition. The final prediction is based on the sum of features at different pyramid scales.

increases, features of the local parts can be more concentrated on the finer discriminative information in each stripe to filter information on the other stripes. Note that these local regions are not necessary to be semantic partitions but a piece of equally divided stripe on the original images.

- We propose a refine mechanism to HPN to refine the uniform partition. We observe outliers in each part under uniform partition. In fact, these outliers are closer to the content of other parts, which means that there are inconsistencies within some parts. Motivated by the fact that within each partitioned part the contents should be consistent, we refine the uniform partition by relocating those outliers to their closest parts, thereby enhancing internal consistency.
- With the proposed induced training strategy, the convergence is speeded up and the problem of limited training data is overcome. The final prediction is generated based on the confidence guided summation of features at different pyramid scales.
- To demonstrate the superiority of our proposed method, we employ experiments on lab-controlled facial expression dataset (Oulu-CASIA) and real-world facial expression dataset (RAF-DB). Our facial expression recognition solution achieves state-of-the-art results on Oulu-CASIA and RAF-DB with accuracies of 90.28% and 86.83%, respectively.

II. RELATED WORK

In this section, the recent related works are reviewed. The deep learning based facial expression recognition methods are discussed firstly. Then we roughly classify the most related works of our proposed method into global based methods and part based methods. The general approach of those methods, their advantages and disadvantages are analyzed in detail.

A. DEEP FACIAL EXPRESSION RECOGNITION

Deep learning methods [12], [19]–[23] for facial expression recognition have achieved great success in the past few years. Yang *et al.* [12] proposed a de-expression residue learning with a generative model, based on the observation that a facial expression can be regarded as a combination of neutral face image and the expressive component. To reduce the identity influence, Cai *et al.* [19] proposed an identity-free generative adversarial network. It holds expression and generates an average identity face image. Zhang *et al.* [20] considered the pose variation and leveraged adversarial autoencoder to augment the training set with expression and poses. However, these methods mainly focused on datasets captured in controlled environment. Thus, those models generalize poorly for uncontrollable variations. In the work of Zeng *et al.* [21], each image is predicted with multiple pseudo labels and a model is learned to fit the latent truth from these inconsistent labels. Acharya *et al.* [23] explored a covariance pooling layer to capture the distortions in regional facial features and temporal evolution of per-frame features. Although the aforementioned approaches achieve good performance on data in the wild, facial expression recognition is still challenging due to the existence of specific local variations.

B. GLOBAL-BASED METHODS

Existing non part-based methods that address expression recognition can be divided into two categories. The first category tries to boost the performance by designing loss functions. Cai *et al.* [8] extend center loss [24] to island

loss which can reduce intra-class variations and augment inter-class differences. Liu *et al.* [9] extend triplet loss [25] to *(N+M)*-triplet clusters loss by incorporating N negative sets with same identity and M examples with same expression. Li and Deng [10] propose a locality-preserving loss which preserve the local proximity by minimizing the distance to K-nearest neighbors within the same class. The second category attempt to make the network disentangle the identity and the expression. Zhang *et al.* [11] propose an expression-identity fusion network. It jointly learns identity-related features and expression-related features via two branches from a input expression. Yang *et al.* [12] use a de-expression residua learning expressive component in a generative model. In Identity-Adaptive Generation (IA-gen) method [13], any given input facial image is transferred to six expressive images of the same subject using six conditional generative adversarial networks. Then, expression classification is performed by comparing the input image with the six generated expressive images.

C. PART-BASED METHODS

Recently, there are some works starting to investigate the part based cues. They can also be divided into two categories according to how they acquire local features. The first category locates the ROI and extract their local features. Patch-gated Convolutional Neural Network [14] decomposes a face into different patches based on facial landmark. A patch-gated unit explicitly predicts the facial occlusion likelihood. In [16], the local features are extracted from crucial regions located by the attention maps. The second category extract local features through cropped facial patches. Liu *et al.* [17] model a boosted deep belief network to classify different expressions. They divide expressional images into patches and select high discriminative patches to train a strong classifier. Xie and Hu [18] propose a convolutional neural network with two branches. One branch extract local features from uniform image patches while the other extract holistic features from the whole expressional image.

III. METHOD

IV. HORIZONTAL PYRAMID NETWORK

In this section we describe the structure of Horizontal Pyramid Network (HPN) as shown in Figure 2. The input image is fed into a backbone network to extract the feature maps. Then we use Pyramidal Average Pooling (PAP) to obtain the feature vector of each local and global spatial stripes. Each feature vector is input into a non-share fully connected layer and followed by a softmax to make the classification. The details are given in the following.

A. NETWORK ARCHITECTURE

When the deep feature map is divided into spatial stripes of different numbers, the size of spatial stripes also changes. Different sizes of spatial stripes contain different amounts of information. As the number of partitions increase, the information in each spatial stripe will be limited. Deep convolutional neural network (DCNN) can capture approximate response preferences on the main body from the whole image. It is also able to capture more fine-grained saliency for local features extracted from smaller local regions. Specifically, when we shrink the representation region and train it as a classification task for learning local features, the supervised signal forces the features to be correctly classified as target expressions. It will drive the learning process to try to explore useful fine-grained details in limited information. By changing the number of spatial stripes, we can obtain local feature representations with multiple granularities. Inspired by the above observations and analysis, we propose HPN to combine global and multi-granularity local feature which is a more powerful facial expression representation.

The architecture of HPN is shown in Figure 2. We choose a variant of Densenet [26] as the backbone. It consists of 3 dense blocks and 2 transition layers. The dense block contains 6, 12 and 24 dense layers, respectively. The input image is fed into the backbone network to extract the feature maps.

According to facial physical structure of a human face, we divide the feature maps into multiple strips horizontally from top to down. We use PAP to obtain vectors of fixed length for facial parts with three kinds of horizontal pyramid scales. These vectors are further fed into fully-connected layer. In this way, the discriminative ability of facial parts can be captured from global to local.

The proposed HPN contains three branches. The first branch learns the feature representations without any partition information, so it obtains the global representation of the whole face. Feature maps in the second branch are uniformly split into two stripes horizontally. It can be seen as a process of obtaining local representation of upper half face and bottom half face. The third branch splits the feature maps into three uniform stripes. It obtains a finer local representation of face, which can be regarded as the local region of eyebrow-eye, nose and mouth. Formally, denote the feature maps extracted by backbone as F. We adopt 3 pyramid scales within the PAP and F is sliced into several spatial stripes horizontally and equally according to different scales. Specifically, assume each spatial stripe as $F_{i,j}$. i, j stand for the index of scale and stripe in each scale. For instance, $F_{3,2}$ means the second strip in third pooling scale. Then, we pool each spatial strip $F_{i,j}$ by horizontal average pooling to generate feature vector $V_{i,j}$. Figure 3 shows the operation details of horizontal average pooling from a plane perspective. Where f is the column vector in the same stripe of F. For example, $V_{3,1}$ is equal to the average of $f_{1,1}, f_{1,2}, \ldots f_{2,5}, f_{2,6}$, that is, $V_{3,1} = 1/12(f_{1,1} + f_{1,2} + \ldots + f_{2,5} + f_{2,6})$.

Besides, we employ the sum fusion to aggregate all the global and local feature vector. It can perfect the comprehensiveness for learned features. The resulting hybrid feature is marked as $V_{0,0}$. After that, we can obtain vectors of fixed length for facial parts at different horizontal pyramid scales. Finally, each feature vector V is input into a independent classifier to predict the expression of the input. The classifier is implemented with a fully-connected (FC) layer

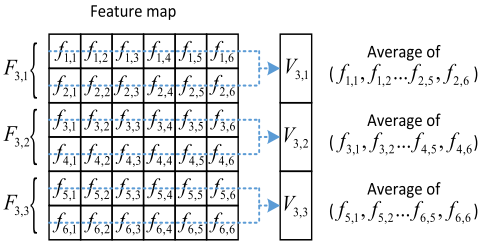


FIGURE 3. Schematic diagram of horizontal average pooling operation.

and a softmax layer. Each classifier is trained by backpropagating the recognition loss independently. Training multiple classifiers adds more supervision and acts as a regularizer to alleviate the overfitting issue. Besides, the discriminative description covers more fine-grained partial features with the increasing pyramid scales. During testing phases, we make the prediction based on the hybrid feature $V_{0,0}$.

B. LOSS FUNCTION

We use fully connected layer as the classifier, each feature vector $V_{i,j}$ is fed into a corresponding classifier $FC_{i,j}$ and following a softmax to classify the expression. During training, the output of given image I is a set of predictions $\hat{q}_{i,j}$. Each $\hat{q}_{i,j}$ can be formulated as

$$\hat{q}_{i,j} = \arg\max_{c \in N} \frac{exp((W^c_{i,j})^T V_{i,j}(I))}{\sum_{n=1}^{N} exp((W^n_{i,j})^T V_{i,j}(I))} \tag{1}$$

where the N is the total number of expression classes, $W_{i,j}$ is the weights of $FC_{i,j}$, q is the ground truth expression of input image I. The loss function is sum of Cross Entropy loss of each output $\hat{q}_{i,j}$.

$$Loss = \sum_{m=1}^{M} \sum_{i,j} CE(\hat{q}^m_{i,j}, q^m) \tag{2}$$

where M is the size of mini-batch, CE is the Cross Entropy loss.

V. REFINED HORIZONTAL PYRAMID NETWORK

Uniform partition in HPN is simple, effective, and yet to be improved. This section explains the inconsistency phenomenon in uniform partition and proposes the refinement mechanism as a remedy to enhance within-stripe consistency. We call this model as Refined Horizontal Pyramid Network (RHPN).

A. WITHIN-STRIPE INCONSISTENCY

After the output feature map F is uniformly partitioned, we further emphasize the content consistency within different horizontal pyramid scales. Specifically, the within-stripe consistency can be interpreted as: column vector f in the same stripe of F should be similar to each other, but be discriminative to column vectors in other stripes. Otherwise, the phenomenon of within-stripe inconsistency occurs. It means the stripes are not partitioned properly. Notice that, we do not consider the within-stripe inconsistency in the first branch because it does not divide the feature map. The following analyses are only for the second and third branches in Figure 2.

After training HPN to convergence, for the second and third branches, we compare the similarity between each column vector f and the average-pooled feature vector $V_{i,j}$ of each stripe by measuring the cosine distance. If f is closest to $V_{i,j}$, we infer f to be closest to j strip in i pooling scale. In this way, we find the closest stripe of each f, as shown in Figure 4. Each column vector is represented by a small rectangle and is drawn with the color of its closest stripe. From the above comparison, we find most of the column vectors in the same horizontal stripe are clustered together even with no explicit constraint. However, there are many outliers in the horizontal stripes. The existence of these outliers indicates that they are inherently more consistent with the column vectors in the other stripe. Indeed, they disturb our final expression recognition in some extent.

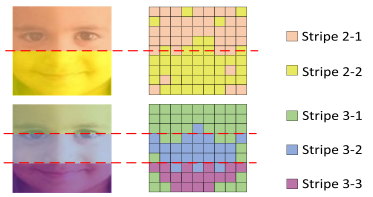


FIGURE 4. Visualization of within-stripe inconsistency.

B. REFINEMENT MECHANISM

We add a refinement mechanism to the second and third branches in HPN to correct within-stripe inconsistency. Our motivation is to assign every column vector to each stripe based on the similarities so that we can relocate outliers.

To this end, we need to dynamically classify every column vector f in F. Based on the learned F, we embed two different Refinement Nets into the second and third branches of HPN, as shown in Figure 5. Each Refinement Net is a classifier consisting of a linear layer and followed by softmax activation which is defined as follows:

$$P(S_i \mid f) = \frac{exp(\Theta_i^T f)}{\sum_{j=1}^{p} exp(\Theta_j^T f)} \tag{3}$$

where P is the predicted probability of f belonging to stripe S_i in each branch, p is the number of pre-defined stripes (i.e., $p = 2$ in the second branch, $p = 3$ in the third branch), and Θ is the trainable weights of the Refinement Net.

Given a column vector f in F and the predicted probability of f belonging to stripe S_i, we assign f to stripe S_i with confidence $P(S_i \mid f)$. Correspondingly, each stripe $S_i (i = 1, \ldots, p)$ is sampled from all column vectors f with $P(S_i \mid f)$ as the sampling weight, i.e.,

$$S_i = \left\{ P(S_i \mid f) \times f, \forall f \in \bar{F} \right\} \tag{4}$$

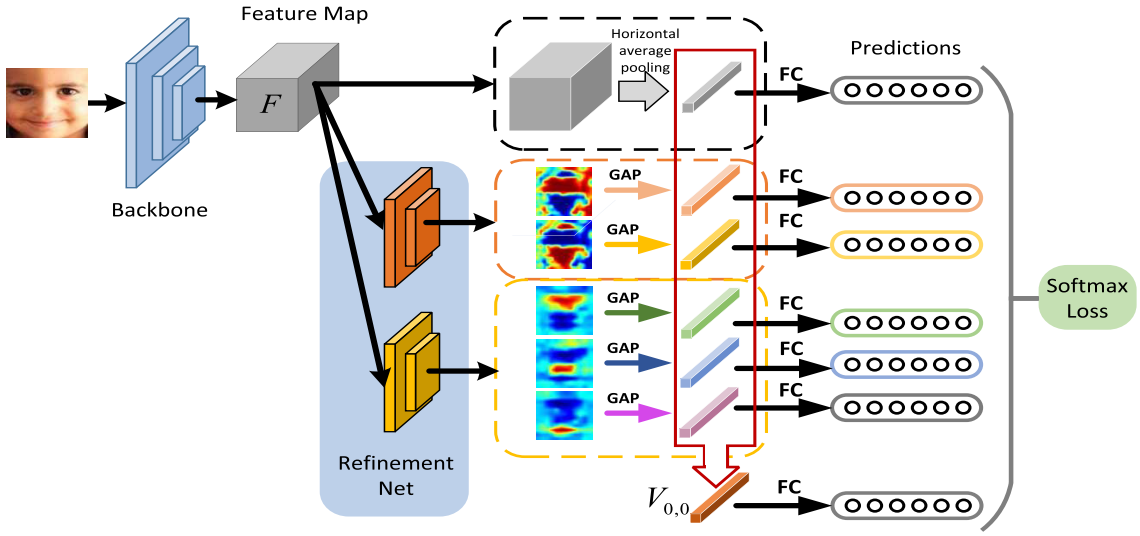


FIGURE 5. Refined Horizontal Pyramid Network. Layers before *F* and after *V* are omitted as they remain unchanged compared with Figure 2. The Refinement Net is added to the second and third branch to enhance within-stripe consistency. GAP denotes global average pooling.

where $\bar{F}$ is the complete set of column vectors in feature map F, $\{\cdot\}$ denotes the sampling operation to form an aggregate.

After the employment of above refinement mechanism, the outliers originated from the uniform partition will be relocated. With the embedding of Refinement Net, HPN is further reshaped into Figure 5. Refinement Net along with the following sampling operation replaces the original average pooling. The structure of all other layers remains exactly the same as in Figure 2.

C. INDUCED TRAINING FOR REFINEMENT NET

Due to the lack of explicit supervisory information for learning Θ of the Refinement Net in Eq. 3, we design an induced training strategy. Specifically, first, a standard HPN is trained to convergence with F equally partitioned. We use this model to induce the training of the Refinement Net in the following steps. Second, we remove the original average pooling layers after F and append a p-category (i.e., $p = 2$ in the second branch of HPN, $p = 3$ in the third branch of HPN) Refinement Net on F. New parts are sampled from F according to the prediction of the Refinement Net. Third, we froze all the already learned layers in HPN, only make the Refinement Net trainable. Then we retrain the model on training set. In this condition, the model still expects the tensor F to be equally partitioned, otherwise it will predict incorrect about the expressions of training images. So, the third step penalizes the Refinement Net until it conducts partition close to the original uniform partition, whereas the part classifier is prone to categorize inherently similar column vectors into a same stripe. A state of balance will be reached as a result of the third step. Finally, all the layers are allowed to be updated. HPN along with the Refinement Net are fine-tuned for overall optimization.

D. CONFIDENCE GUIDED PREDICTION

In the proposed RHPN, the final prediction is based on the summation of features at different pyramid scales. However, the outputs of softmax of every classifier $FC_{i,j}$ describe the confidence of feature vector $V_{i,j}$ supporting the predicted expression class. In other words, the output probability of softmax provide us the useful estimation of classification uncertainty. Uncertainty estimation has a long history in neural networks, starting with Bayesian neural networks [27]. It aims at producing estimates that are function of network parameters and the input image and thus reason about the current observations. Motivated by the work [28], we define the final prediction as a kind of confidence guided way. For simplicity, let us define a confidence value s_i, j.

$$s_{i,j} = \max \frac{exp((W_{i,j}^{c})^T V_{i,j}(I))}{\sum_{n=1}^{N} exp((W_{i,j}^{n})^T V_{i,j}(I))} \tag{5}$$

The final prediction is calculated as the softmax of the weighted sum of all feature vectors.

$$\hat{q}_{0,0} = \arg\max_{c \in N} \frac{exp(s_{i,j}(W_{i,j}^{c})^T V_{i,j}(I))}{\sum_{n=1}^{N} exp(s_{i,j}(W_{i,j}^{n})^T V_{i,j}(I))} \tag{6}$$

The confidence guided prediction is important for avoiding the finagle of prediction based on the separate feature vector. Due to changes in posture or personal habits, the representation of partial region may be very different from the overall representation of expression of the current image. At this time, the classification of expression based on the feature vector extracted from the local region will no longer be reliable. Therefore, even if our model is able to rely on such a feature vector to make a prediction, its confidence is very low. When we make the final classification, the confidence weighted summation will reduce the influence of the

unconfidence feature vector on the final classification result. Therefore, the confidence-guided prediction makes the whole model robust to local region disturbances.

VI. EXPERIMENTS

In this section, we briefly introduce the datasets and implementation details. Then we report the performances compared with some recent state of the arts. Finally we evaluate multiple basic instantiations of the proposed method to analyze the effects of those core factors.

A. DATASET AND PREPROCESSING

Most of our experiments are conducted on the Oulu-CASIA dataset [29]. It is a lab-controlled dataset which contains 480 facial expression sequences collected from 80 different subjects. Each sequence begins with a neutral expression and ends with a peak expression. All sequences have been labeled with six expressions: anger, disgust, fear, happiness, sadness or surprise. As a general procedure [8], [11], [13], [16], [31], [32], the last three frames of each sequence are used for training and test. Thus, Oulu-CASIA contains 1440 images for our experiments. Additionally, we also conduct experiments on the Real-world Affective Face Database (RAF-DB) [10]. It is a real-world dataset that contains 29,672 highly diverse facial images downloaded from the Internet. Images labeled with seven basic expressions (surprise, fear, disgust, happiness, sadness, anger and neutral) are used in our experiment, including 12,271 images for training and 3,068 images for test.

B. IMPLEMENT DETAILS

Face alignment is conducted based on the facial landmarks detected with Supervised Descent Method (SDM) [33]. The detected face are cropped, resized and converted to 60 $\times$ 60 gray scale images. We ignore extra alignment method in RAF-DB because face images have already been aligned. To avoid over-fitting, we utilize conventional data augmentation method to generate more training data, where each image is rotated by degree $\{-15°, -12°, -9°, -6°, -3°, 0°, 3°, 6°, 9°, 12°, 15°\}$ respectively. Five 48 $\times$ 48 patches are cropped out from five locations of each image (center and four corners, respectively), then each patch is flipped horizontally, thus resulting in an augmented dataset which is 110 times larger than the original one. During testing, a single center crop with size 48 $\times$ 48 is used for testing. Neither the rotation nor the flipping operation is used. Since the resolution of input is 48×48, the resolution of the feature map F obtained through the backbone network is 12 $\times$ 12. Thus, the resolution of each strip in the second and third branches are 6 $\times$ 6 and 4 $\times$ 4, respectively. Because Oulu-CASIA does not provide specified training and test sets, we employ the most popular 10-fold validation strategy as in the previous methods [8], [11], [16], [31], [32], [34]. The dataset is split into ten groups without subject overlapping between the groups. For each run, nine groups are used for training and the remaining is used for test. The results are the average of 10 runs. For the experiments on the RAF-DB database, we use their official split for training and test.

For HPN, the backbone is pre-trained on ImageNet [30]. We train the backbone and FC layers for 50 epochs with initial learning rate 0.01 and 0.1 respectively and decayed to 0.1$\times$ after 30 epochs. When employing refinement mechanism, we append another 10 epochs on the Refinement Net with learning rate 0.01. Finally, HPN along with the Refinement Net are fine-tuned with new learning rate 0.001 for the backbone and 0.01 for the remainder. The learning rate decayed to 0.1$\times$ after 20 epochs. The stochastic gradient descent (SGD) with 0.9 momentum is implemented in each mini-batch.

C. EXPRESSION RECOGNITION RESULTS

To compare the performance of the proposed method with others, Table 1 and Table 2 compare it with other competitive approaches through two indicators, namely feature (dynamic feature or static feature) and average recognition accuracy (calculated as percentage) on the Oulu-CASIA and RAF-DB datasets. To evaluate the overall performance, the confusion matrices on two datasets are illustrated in Figure 6 and Figure 7, respectively.

TABLE 1. Performance comparison on the Oulu-CASIA.

Method	Feature	Accuracy
HOG 3D [35]	Dynamic	70.63
STM-ExpLet [36]	Dynamic	74.59
IL-CNN [8]	Static	77.29
DTAGN [37]	Dynamic	81.86
LOMo [38]	Dynamic	82.10
PPDN [40]	Static	84.59
EIFN [11]	Static	85.21
PHRNN-MSCNN [39]	Dynamic	86.25
GCNet [41]	Static	86.39
FN2EN [34]	Static	87.71
DeRL [12]	Static	88.00
IDFERM [32]	Static	88.25
WS-LGAN [16]	Static	88.26
IA-gen [13]	Static	88.92
DE-GAN [31]	Static	89.17
HPN	Static	88.63
RHPN	Static	90.33

TABLE 2. Performance comparison on the RAF-DB.

Method	Feature	Accuracy
FSN [44]	Static	81.10
baseDCNN [10]	Static	82.86
PG-CNN [14]	Static	83.27
Center Loss [10]	Static	83.68
DLP-CNN [10]	Static	84.13
PAT-ResNet(gender,race) [45]	Static	84.19
Lin et al. [46]	Static	84.68
gACNN [15]	Static	85.07
WS-LGAN [16]	Static	85.07
APM-VGG [47]	Static	85.17
HPN	Static	83.53
RHPN	Static	86.86

1) EVALUATION ON OULU-CASIA

In Table 1, the proposed HPN achieves an average recognition accuracy of 88.54% on the Oulu-CASIA dataset. It surpasses most of the previous methods. The performance is boosted by RHPN. It exceeds HPN by +1.74% on the Oulu-CASIA datasets. RHPN yields the higher accuracy than all the state-of-the-arts, which demonstrates the effectiveness of the added refinement mechanism. Figure 6 shows that RHPN performs well when recognizing happiness and surprise, which reaches the accuracy of 97%. Disgust and anger are seriously confused. The main reason is they act similarly in some facial action units in FACS, such as AU10 (Upper Lip Raiser), AU17 (Chin Raiser), AU25 (Lips Part) and AU26 (Jaw Drop) [42], [43].

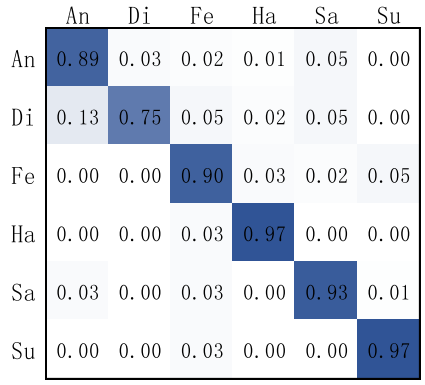


FIGURE 6. Confusion matrices on the Oulu-CASIA.

2) EVALUATION ON RAF-DB

Table 2 shows the average recognition accuracy on the RAF-DB. Some papers report performance as an average of diagonal values of confusion matrix. We convert them to regular accuracy for fair comparison. Unlike the Oulu-CASIA, the RAF-DB is more challenging as it has a wider variety of pose and a larger range of ages. The data is closer to the natural scene. HPN achieves an average recognition accuracy of 83.44%. Compared with other methods, this result is not impressive. This is mainly because the face pose of the image in RAF-DB varies greatly. Relying on uniform partition only, HPN will introduce a large number of outliers, which hinders the performance. After we relocate the outliers originated from the uniform partition through the Refinement Net, the recognition accuracy of RHPN further surpasses HPN by a gain of +3.39%. Compared with other state-of-the-art algorithms, our proposed RHPN achieves substantial improvement. It is worth noting that PG-CNN and gACNN leverage external facial landmarks to localize the discriminative regions. WS-LGAN leverage external facial attribute dataset and attention map to localize the discriminative regions. The result proves that our method is also robust to real-world facial expression dataset. Figure 7 shows that the highest accuracy is obtained when recognizing happiness, which reaches to 94%. However, the performance on anger, disgust and fear are poor. This is mainly due to the lack of data. In RAF-DB the samples of anger, disgust and fear are far less than others.

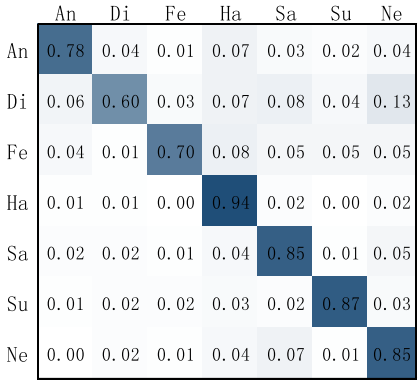


FIGURE 7. Confusion matrices on the RAF-DB.

D. ABLATION STUDIES

To verify the effectiveness of components, we conduct several ablation studies on Oulu-CASIA, including influence of backbone, different number of pyramid scales, w/ and w/o refinement mechanism, induced training and different predicting strategy. Note that all unrelated settings are the same as HPN and RHPN implementation.

1) INFLUENCE OF BACKBONE

In the architecture of our network, we choose a variant of Densenet [26] as the backbone. Indeed, we also have chosen several basic CNNs as our backbone. For simplicity, we attach the structure of HPN to every backbone. The different networks are initialized based on their pretrained model from ImageNet dataset. The performances are shown in Table 3. We have verified VGG-16 [48] and two kinds of ResNet [49]. We can see that the best performance is achieved by Densenet. It is worth note that one can boost the performance of our proposed method by implementing a more powerful backbone.

TABLE 3. Evaluation of backbone.

Backbone	Accuracy(%)
VGG-16	76.92
ResNet-50	80.53
ResNet-101	85.31
DenseNet	88.63

2) NUMBER OF PYRAMID SCALES

Table 4 shows the results of HPN and RHPN with different pyramid scales. We can find that HPN and RHPN reach the best performance with three pyramid scales. When the pyramid scale is sets to 1, it is equivalent to global pooling. With the increasing of pyramid scales' numbers, the accuracy of HPN and RHPN are significant improved. This is reasonable as smaller partition will force the model to extract

TABLE 4. Performance comparison of the proposed method with different pyramid scales.

pyramid scale	spatial stripes	w/o ref (%)	w/ ref (%)
1	1	86.18	–
2	1,2	86.80	87.50
3	1,2,3	88.54	90.28

TABLE 5. Evaluation of effectiveness of pyramid structure.

Model	Accuracy(%)
w/o pyramid	87.43
RHPN	90.28

more discriminative fine-grained features. Besides, compared with HPN, the recognition accuracy is improved by adding refinement mechanism at any pyramid scales, which further proves the superiority of RHPN. In fact, we also try more dense pyramid scales, such as 4, 5 and 6. However, more pyramid scales will bring additional computational cost but no obvious improvement can be observed. It is because the parts at these pyramid scales are too small to learn discriminative information. Therefore, we finally adopt three pyramid scales.

3) EFFECTIVENESS OF PYRAMID STRUCTURE

Previous analysis shows that the HPN and RHPN reaches the best performance with three pyramid scales, which has up to 3 partial stripes on the feature map. In order to verify the effectiveness of pyramid structure, we remove other branches and just preserve the branch with 3 partial stripes. From Table 5, we can observe that with our pyramid structure, the performance is further improved by a large margin. This is due to the pyramid structure can formulate partial features from coarse to fine and combine both global and multi-grained local features. It greatly increases the diversity and discriminative ability of the features.

4) INDUCED TRAINING

In the training phrase of network, we can also conduct a general training strategy. We directly impose the probability to the feature vector in the refinement net, and then fine-tune all the parameters in the RHPN network on the Oulu-CASIA dataset. Table 6 shows the results of HPN and RHPN with different training strategies. We can find that RHPN reach the best performance with our proposed induced training strategy. It is worth note that the RHPN with general training strategy has only a slight improvement over HPN. This indicates that in the HPN training phase, although column vectors exist within-stripe inconsistency, the stripes categories are maintained. But in general training strategy, the effect of refinement net is hardly displayed because of the complete loss of an effective classification prior.

TABLE 6. Evaluation of training strategies.

Training strategy	Accuracy(%)
HPN	88.63
RHPN General training	88.65
RHPN Induced training	90.33

5) PREDICTING STRATEGY

In order to prove the effect of our proposed confidence guided prediction, we also implement other predicting strategies such as voting and summation. Table 7 shows the results of predicting strategies. The first row is the result of voting on the classification based on the features at different pyramid scales. The second row is the result of classification on the summation of features at different pyramid scales. The last row is the result of our proposed confidence guided prediction. We can find that voting has a unsatisfied performance. The summation seems better than voting. Our confidence guided prediction achieves the best performance.

TABLE 7. Evaluation of predicting strategies.

Predicting strategy	Accuracy(%)
voting	85.34
summation	90.28
confidence	90.33

6) VISUALIZATION

A visualization of the refined feature response with multiple granularity on the Oulu-CASIA is illustrated in Figure 8. Compare with the feature responses in the second branch of RHPN (Figure 8(b)), feature responses in the third branch (Figure 8(c)) are more scattered on facial parts, but some pivotal semantic information is preferred. It can filter most of the complex background which contains interfere information of facial expression. In addition, some outliers in each feature stripe in the third branch have also been adjusted. Compare

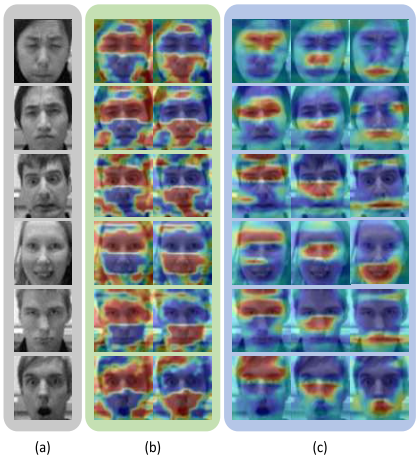


FIGURE 8. Visualization of the refined feature response. (a) original images. (b) response maps from the second branch. (c) response maps from the third branch.

with the third branch of RHPN, the second branch is a more coarse partition. Therefore, each stripe in the second branch contains more outliers. From Figure 8(b) we can see that many outliers are restored. The feature responses show that some features of forehead are more similar to those of bottom half face, while some features of cheek are more similar to those of upper half face.

VII. CONCLUSION

This paper makes two main contributions to facial expression recognition. Firstly, we propose a Horizontal Pyramid Network (HPN). It employs a uniform partial feature representation at different horizontal pyramid scales. HPN can exploit various partial information of facial image without of any part annotation, which successfully enhances the diversity and discriminative ability of the features. Furthermore, we add a refinement mechanism to HPN. It reinforces the within-stripe consistency in each stripe in HPN. With induced training strategy and confidence guided prediction, the refinement mechanism requires no part labeling information and further improves the performance considerably. Extensive experiments on several datasets demonstrate the effectiveness of our proposed method.

REFERENCES

[1] A. Ryan, J. F. Cohn, S. Lucey, J. Saragih, P. Lucey, F. De la Torre, and A. Rossi, "Automated facial expression recognition system," in *Proc. Int. Carnahan Conf. Secur. Technol.*, Oct. 2009, pp. 172–177.

[2] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image Vis. Comput.*, vol. 27, no. 12, pp. 1743–1759, Nov. 2009.

[3] Q. Wang, K. Jia, and P. Liu, "Design and implementation of remote facial expression recognition surveillance system based on PCA and KNN algorithms," in *Proc. Int. Conf. Intell. Inf. Hiding Multimedia Signal Process.*, Sep. 2015, pp. 314–317.

[4] E. Vural, M. Çetin, A. Erçil, G. Littlewort, M. Bartlett, and J. Movellan, "Automated drowsiness detection for improved driving safety," in *Proc. Int. Conf. Automat. Technol.*, 2008, pp. 1–15.

[5] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[6] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 6, pp. 1113–1133, Jun. 2015.

[7] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affect. Comput.*, early access, Mar. 17, 2020, doi: 10.1109/TAFFC.2020.2981446.

[8] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly, and Y. Tong, "Island loss for learning discriminative features in facial expression recognition," in *Proc. 13th IEEE Int. Conf. Automat. Face Gesture Recognit.*, May 2018, pp. 302–309.

[9] X. Liu, B. V. K. V. Kumar, J. You, and P. Jia, "Adaptive deep metric learning for identity-aware facial expression recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 20–29.

[10] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 356–370, Jan. 2019.

[11] H. Zhang, W. Su, and Z. Wang, "Expression-identity fusion network for facial expression recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2019, pp. 2122–2126.

[12] H. Yang, U. Ciftci, and L. Yin, "Facial expression recognition by de-expression residue learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2168–2177.

[13] H. Yang, Z. Zhang, and L. Yin, "Identity-adaptive facial expression recognition through expression regeneration using conditional generative adversarial networks," in *Proc. 13th IEEE Int. Conf. Automat. Face Gesture Recognit.*, May 2018, pp. 294–301.

[14] Y. Li, J. Zeng, S. Shan, and X. Chen, "Patch-gated CNN for occlusion-aware facial expression recognition," in *Proc. 24th Int. Conf. Pattern Recognit.*, Aug. 2018, pp. 2209–2214.

[15] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using CNN with attention mechanism," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2439–2450, 2018.

[16] H. Zhang, W. Su, and Z. Wang, "Weakly supervised local-global attention network for facial expression recognition," *IEEE Access*, vol. 8, pp. 37976–37987, 2020.

[17] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1805–1812.

[18] S. Xie and H. Hu, "Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks," *IEEE Trans. Multimedia*, vol. 21, no. 1, pp. 211–220, Jan. 2019.

[19] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly, and Y. Tong, "Identity-free facial expression recognition using conditional generative adversarial network," 2019, *arXiv:1903.08051*. [Online]. Available: http://arxiv.org/abs/1903.08051

[20] F. Zhang, T. Zhang, Q. Mao, and C. Xu, "Joint pose and expression modeling for facial expression recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3359–3368.

[21] J. Zeng, S. Shan, and X. Chen, "Facial expression recognition with inconsistently annotated datasets," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 222–237.

[22] Z. Luo, J. Hu, and W. Deng, "Local subclass constraint for facial expression recognition in the wild," in *Proc. 24th Int. Conf. Pattern Recognit.*, Aug. 2018, pp. 3132–3137.

[23] D. Acharya, Z. Huang, D. P. Paudel, and L. Van Gool, "Covariance pooling for facial expression recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 367–374.

[24] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 499–515.

[25] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 815–823.

[26] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4700–4708.

[27] M. Welling and Y. W. Teh, "Bayesian learning via stochastic gradient Langevin dynamics," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 681–688.

[28] Y. Shi, X. Yu, K. Sohn, M. Chandraker, and A. K. Jain, "Towards universal representation learning for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 6817–6826.

[29] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen, "Facial expression recognition from near-infrared videos," *Image Vis. Comput.*, vol. 29, no. 9, pp. 607–619, 2011.

[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[31] K. Ali and C. E. Hughes, "Facial expression recognition using disentangled adversarial learning," 2019, *arXiv:1909.13135*. [Online]. Available: http://arxiv.org/abs/1909.13135

[32] X. Liu, B. V. K. V. Kumar, P. Jia, and J. You, "Hard negative generation for identity-disentangled facial expression recognition," *Pattern Recognit.*, vol. 88, pp. 1–12, Apr. 2019.

[33] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 532–539.

[34] H. Ding, S. K. Zhou, and R. Chellappa, "FaceNet2ExpNet: Regularizing a deep face recognition net for expression recognition," in *Proc. 12th IEEE Int. Conf. Automat. Face Gesture Recognit.*, May 2017, pp. 118–126.

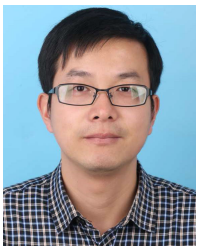
[35] A. Klaser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *Proc. 19th Brit. Mach. Vis. Conf.*, Sep. 2008, p. 275.

[36] M. Liu, S. Shan, R. Wang, and X. Chen, "Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1749–1756.
[37] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 2983–2991.
[38] K. Sikka, G. Sharma, and M. Bartlett, "LOMo: Latent ordinal model for facial analysis in videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 5580–5589.
[39] K. Zhang, Y. Huang, Y. Du, and L. Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4193–4203, Sep. 2017.
[40] X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, N. Vasconcelos, and S. Yan, "Peak-piloted deep network for facial expression recognition," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 425–442.
[41] Y. Kim, B. Yoo, Y. Kwak, C. Choi, and J. Kim, "Deep generative-contrastive networks for facial expression recognition," 2017, *arXiv:1703.07140*. [Online]. Available: http://arxiv.org/abs/1703.07140
[42] P. Ekman and W. V. Friesen, *Facial Action Coding System: Investigator's Guide*. Consulting Psychologists Press, 1978.
[43] E. Friesen and P. Ekman, "Facial action coding system: A technique for the measurement of facial movement," *Palo Alto*, vol. 3, no. 2, p. 5, 1978.
[44] S. Zhao, H. Cai, H. Liu, J. Zhang, and S. Chen, "Feature selection mechanism in CNNs for facial expression recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2018, p. 317.
[45] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly, and Y. Tong, "Probabilistic attribute tree in convolutional neural networks for facial expression recognition," 2018, *arXiv:1812.07067*. [Online]. Available: http://arxiv.org/abs/1812.07067
[46] F. Lin, R. Hong, W. Zhou, and H. Li, "Facial expression recognition with data augmentation and compact feature learning," in *Proc. 25th IEEE Int. Conf. Image Process.*, Oct. 2018, pp. 1957–1961.
[47] Z. Li, S. Han, A. S. Khan, J. Cai, Z. Meng, J. O'Reilly, and Y. Tong, "Pooling map adaptation in convolutional neural network for facial expression recognition," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2019, pp. 1108–1113.
[48] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

WEN SU was born in 1992. She received the B.E. degree in engineering from the Department of Automation, University of Science and Technology of China, in 2013, and the Ph.D. degree in control science and engineering from the University of Science and Technology of China, in 2018. She currently works with the Virtual Reality Laboratory, Zhejiang Sci-Tech University. Her research interests include image segmentation and depth scene understanding based on deep learning.

HAIFENG ZHANG was born in 1993. He received the B.E. degree from the China University of Geosciences, Beijing, in 2015. He is currently pursuing the Ph.D. degree with the University of Science and Technology of China. His work published in several different journals and conferences. His research interests include face recognition, facial expression recognition, and depth scene understanding based on deep learning.

YUAN SU was born in 2000. She is currently pursuing the B.S. degree with the Xi'an University of Technology. Her research interests include computer vision, deep learning, object detection, and emotion recognition.



JUN YU is currently an Associate Professor with the Department of Automation, University of Science and Technology of China. He has published more than 100 journal articles and conference papers, such as IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, and LANGUAGE PROCESSING, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, *TOMM*, ACL, CVPR, *MM*, SIGGRAPH ASIA, *VR*, AAAI, and IJCAI. His research interests include multimedia computing, multi-modal information synthesis, perception, and cognition. He has received two Best Paper awards from premier conferences, namely ICME, FG, and won 12 Grand Challenge Champions and the First Runner-Up aAward from premier conferences, such as CVPR, MM, ECCV, ICME, and FG.